N° d'ordre NNT : 2024ISAL0111

THESE de DOCTORAT DE L'INSA LYON, membre de l'Université de Lyon

Ecole Doctorale N° 512
INFOMATHS

Spécialité / discipline de doctorat :
Informatique

Soutenue publiquement le 09/12/2024, par :
Pierre-Yves Genest

Unsupervised Open-World Information Extraction From Unstructured and Domain-Specific Document Collections

---

Devant le jury composé de :

Mothe, Josiane	Professeur des Universités, INSPÉ Toulouse Occitanie-Pyrénées, France	Présidente
Gianini, Gabriele	Professeur des Universités, Università degli Studi di Milano-Bicocca, Italie	Rapporteur
Granitzer, Michael	Professeur des Universités, Universität Passau, Allemagne	Rapporteur
Calabretto, Sylvie	Professeur des Universités, INSA Lyon, France	Examinatrice
Egyed-Zsigmond, Előd	Maître de Conférences HDR, INSA Lyon, France	Directeur de thèse
Portier, Pierre-Edouard	Docteur, Caisse d'Epargne Rhône Alpes, France	Invité (co-encadrant)
Lovisetto, Martino	Docteur, Alteca, France	Invité

INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Département de la Formation par la Recherche et des Études Doctorales (**FEDORA**)

Bâtiment Charlotte Perriand 3e étage
17, av. des Arts
69621 Villeurbanne Cédex
fedora@insa-lyon.fr

Référence : TH1164_GENEST Pierre-Yves

L'INSA Lyon a mis en place une procédure de contrôle systématique via un outil de détection de similitudes (logiciel Compilatio). Après le dépôt du manuscrit de thèse, celui-ci est analysé par l'outil. Pour tout taux de similarité supérieur à 10%, le manuscrit est vérifié par l'équipe de FEDORA. Il s'agit notamment d'exclure les auto-citations, à condition qu'elles soient correctement référencées avec citation expresse dans le manuscrit.

Par ce document, il est attesté que ce manuscrit, dans la forme communiquée par la personne doctorante à l'INSA Lyon, satisfait aux exigences de l'Etablissement concernant le taux maximal de similitude admissible.

INSA LYON
Campus LyonTech La Doua
20, avenue Albert Einstein - 69621 Villeurbanne cedex - France
Tél. +33 [0]4 72 43 83 83 - Fax +33 [0]4 72 43 85 00
www.insa-lyon.fr

membre de UNIVERSITÉ DE LYON

Département FEDORA – INSA Lyon - Ecoles Doctorales

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
ED 206 CHIMIE	**CHIMIE DE LYON** https://www.edchimie-lyon.fr Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage secretariat@edchimie-lyon.fr	**M. Stéphane DANIELE** C2P2-CPE LYON-UMR 5265 Bâtiment F308, BP 2077 43 Boulevard du 11 novembre 1918 69616 Villeurbanne directeur@edchimie-lyon.fr
ED 341 E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** http://e2m2.universite-lyon.fr Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.e2m2@univ-lyon1.fr	**Mme Sandrine CHARLES** Université Claude Bernard Lyon 1 UFR Biosciences Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69622 Villeurbanne CEDEX e2m2.codir@listes.univ-lyon1.fr
ED 205 EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** http://ediss.universite-lyon.fr Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.ediss@univ-lyon1.fr	**Mme Sylvie RICARD-BLUM** Laboratoire ICBMS - UMR 5246 CNRS - Université Lyon 1 Bâtiment Raulin - 2ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tél : +33(0)4 72 44 82 32 sylvie.ricard-blum@univ-lyon1.fr
ED 34 EDML	**MATÉRIAUX DE LYON** http://ed34.universite-lyon.fr Sec. : Yann DE ORDENANA Tél : 04.72.18.62.44 yann.de-ordenana@ec-lyon.fr	**M. Stéphane BENAYOUN** Ecole Centrale de Lyon Laboratoire LTDS 36 avenue Guy de Collongue 69134 Ecully CEDEX Tél : 04.72.18.64.37 stephane.benayoun@ec-lyon.fr
ED 160 EEA	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** https://edeea.universite-lyon.fr Sec. : Philomène TRECOURT Bâtiment Direction INSA Lyon Tél : 04.72.43.71.70 secretariat.edeea@insa-lyon.fr	**M. Philippe DELACHARTRE** INSA LYON Laboratoire CREATIS Bâtiment Blaise Pascal, 7 avenue Jean Capelle 69621 Villeurbanne CEDEX Tél : 04.72.43.88.63 philippe.delachartre@insa-lyon.fr
ED 512 INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** http://edinfomaths.universite-lyon.fr Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 infomaths@univ-lyon1.fr	**M. Hamamache KHEDDOUCI** Université Claude Bernard Lyon 1 Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tél : 04.72.44.83.69 direction.infomaths@listes.univ-lyon1.fr
ED 162 MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** http://edmega.universite-lyon.fr Sec. : Philomène TRECOURT Tél : 04.72.43.71.70 Bâtiment Direction INSA Lyon mega@insa-lyon.fr	**M. Etienne PARIZET** INSA Lyon Laboratoire LVA Bâtiment St. Exupéry 25 bis av. Jean Capelle 69621 Villeurbanne CEDEX etienne.parizet@insa-lyon.fr
ED 483 ScSo	**ScSo**[1] https://edsciencessociales.universite-lyon.fr Sec. : Mélina FAVETON Tél : 04.78.69.77.79 melina.faveton@univ-lyon2.fr	**M. Bruno MILLY** (INSA : J.Y. TOUSSAINT) Univ. Lyon 2 Campus Berges du Rhône 18, quai Claude Bernard 69365 LYON CEDEX 07 Bureau BEL 319 bruno.milly@univ-lyon2.fr

[1] ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

枯れた技術の水平思考
Kareta Gijutsu no Suihei Shikō
[Lateral Thinking with Seasoned Technology]

Gunpei Yokoi

Abstract

The exponential growth in data generation has rendered the effective analysis of unstructured textual document collections a critical challenge. This PhD thesis aims to address this challenge by focusing on Information Extraction (IE), which encompasses four essential tasks: Named Entity Recognition (NER), Coreference Resolution (CR), Entity Linking (EL), and Relation Extraction (RE). These tasks collectively enable extracting and structuring knowledge from unformatted documents, facilitating its integration into structured databases for further analytical processes.

Our contributions start with creating Linked-DocRED, the first large-scale, diverse, and manually annotated dataset for document-level IE. This dataset enriches the existing DocRED [1] dataset with high-quality entity linking labels. Additionally, we propose a novel set of metrics for evaluating end-to-end IE models. The evaluation of baseline models on Linked-DocRED highlights the complexities and challenges inherent to document-level IE: cascading errors, long context handling, and information scarcity.

We then introduce PromptORE, an unsupervised and open-world RE model. Adapting the prompt-tuning paradigm, PromptORE achieves relation embedding and clustering without requiring fine-tuning or hyperparameter tuning (a major weakness of previous baselines) and significantly outperforms state-of-the-art models. This method demonstrates the feasibility of extracting semantically coherent relation types in an open-world context.

Further extending our prompt-based approach, we develop CITRUN for unsupervised and open-world NER. By employing contrastive learning with off-domain labeled data, CITRUN improves entity type embeddings, surpassing LLM-based unsupervised NERs, and achieving competitive performance against zero-shot models that are more supervised.

These advancements facilitate meaningful knowledge extraction from unstructured documents, addressing practical, real-world constraints and enhancing the applicability of IE models in industrial contexts.

Keywords information extraction, open named entity recognition, unsupervised named entity recognition, open relation extraction, unsupervised relation extraction, natural language processing.

Résumé

La croissance exponentielle de la production de données a fait de l'analyse de collections de documents textuels non structurés un défi majeur. Cette thèse de doctorat vise à relever ce défi en se concentrant sur l'extraction d'information (IE), qui englobe quatre tâches principales : reconnaissance d'entités nommées (NER), résolution des coréférences (CR), annotation sémantique (EL) et extraction de relations (RE). Ces tâches permettent d'extraire et de structurer des connaissances à partir de documents non formatés, ce qui facilite leur intégration dans des bases de données structurées et leur utilisation par des outils d'analyse de données.

Nos contributions commencent par la création de Linked-DocRED, le premier jeu de données de grande taille, diversifié, et annoté manuellement pour l'IE sur des documents. Pour cela, nous partons du jeu de données DocRED [1], que nous complétons avec des annotations sémantiques de haute qualité. Également, nous proposons un nouvel ensemble de métriques pour évaluer les modèles d'extraction d'information. L'évaluation de baselines sur Linked-DocRED met en évidence les complexités et les défis inhérents à l'IE sur des documents : erreurs en cascade, traitement de longs contextes et rareté de l'information.

Nous présentons ensuite PromptORE, un modèle d'extraction de relations non supervisé et en monde ouvert. En adaptant le paradigme du prompt-tuning, PromptORE réalise la représentation et le clustering de relations sans nécessiter d'entraînement ni d'ajustement d'hyperparamètres (une faiblesse majeure des baselines précédentes) et surpasse de manière significative les modèles de l'état de l'art. Cette méthode démontre la faisabilité de l'extraction de relations sémantiquement cohérentes dans un contexte de monde ouvert.

En généralisant notre approche basée sur les prompts, nous développons CITRUN, un NER non supervisé et fonctionnant en monde ouvert. En utilisant l'apprentissage contrastif avec des données étiquetées hors domaine, CITRUN améliore la représentation des types d'entités, surpassant les NERs non supervisés basés sur des LLMs, et atteignant des performances compétitives par rapport aux modèles zero-shot qui sont plus supervisés.

Ces avancées facilitent l'extraction de connaissances à partir de documents non structurés, tout en tenant compte des contraintes pratiques du monde réel et en améliorant l'applicabilité des modèles d'IE dans des contextes industriels.

Mots-Clés extraction d'information, reconnaissance d'entités nommées en monde ouvert, reconnaissance d'entités nommées non-supervisée, extraction de relations en monde ouvert, extraction de relations non supervisée, traitement automatique des langues.

Acknowledgements

First, I would like to express my sincere gratitude to my academic and industrial supervisors, Előd Egyed-Zsigmond, Pierre-Edouard Portier, Martino Lovisetto, and Laurent-Walter Goix. I am particularly thankful for your unfailing support, for your many relevant scientific critiques, comments, and directions, and for the continued three-year-long weekly meetings that are the main source of the success of my research work. Előd, thank you for being the central and stable point of my PhD in an ever-changing environment, for diligently handling administrative nonsense, and for so many details, which would be too numerous to list. Pierre-Edouard, I cannot thank you enough for your precise, relevant, and profound guidance, which has been of immense value for exploring research directions and publishing our contributions. Martino, thank you for your patience and for providing me with a healthy working environment at Alteca. Laurent-Walter, thank you for creating a genuine R&D department at Alteca, where we study and develop cutting-edge NLP and AI technologies, and for initializing the academic cooperation that led to my PhD.

I am also grateful to Alteca for giving me the opportunity to pursue my PhD. Speaking of Alteca, I want to thank my colleagues and interns, Yasser, Nour, Quentin, Richard, Mathis, Samuel, Yann, Dhouha, Alexandre, and Juliette, for your knowledge and expertise in your respective areas.

My sincere thanks to all my fellow PhD students, Johan, Paul, Baptiste, Nawel, Matthieu, Yacine, Silvia, Julien, Aghiles, and Brandon, for the captivating scientific discussions we had, but also for the moments spent together having fun. I am also grateful to the members of IRIXYS and the DRIM team for the valuable advice you gave me during workshops and team meetings.

Pour finir, je n'ai pas de mots pour exprimer ma gratitude et ma reconnaissance envers vous, papa, maman, Jean-Samuel. Que dire sinon vous remercier du fond du cœur pour votre soutien indéfectible au quotidien et plus encore pendant les périodes stressantes, pour vos nombreux encouragements, et pour tous ces à-côtés qui ne sont pas scientifiques mais pourtant si essentiels.

Contents

List of Figures

List of Tables

Glossary

AMI Adjusted Mutual Information [2].

ARI Adjusted Rand Index [3, 4].

BIO BIO sequence labeling. Classification convention for Named Entity Recognition [5]. See section V.3.1.

CNN Convolutional Neural Network.

CR Coreference Resolution.

EL Entity Linking.

EncLM Encoder-Only Language Model, for instance, BERT [6].

ER Embedding Refinement. See section V.3.2.3.

ET Entity Typing. See section V.3.2.

FN False Negatives.

FP False Positives.

GCN Graph Convolutional Network [7].

GMM Gaussian Mixture Model.

GNN Graph Neural Network [8].

IE Information Extraction.

KG Knowledge Graph.

LLM Large Language Model.

LSTM Long-Short Term Memory [9].

MD Mention Detection. See section V.3.1.

NER Named Entity Recognition.

RE Relation Extraction.

TN True Negatives.

TP True Positives.

Mathematical Notations

When possible, they follow the ISO-80000-2 standard [10].

α, β, x, y Scalars. In lower-case and normal font.

$\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}, \boldsymbol{y}$ Vectors, lists, or tuples. In lower-case and bold font.

$\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{X}, \boldsymbol{Y}$ Tensors or matrices. In upper-case and bold font.

$\hat{\alpha}, \hat{\boldsymbol{x}}, \hat{\boldsymbol{Y}}$ Predicted scalars, vectors, or tensors.

$\boldsymbol{h}$ Embedding (vector representation) produced by a language model.

σ Softmax function.

$\{\cdot\}$ Variable substitution. $\{\boldsymbol{m}\}$ is replaced by the text of $\boldsymbol{m}$.

t Token. Word, part of word, or punctuation as defined by SentencePiece [11].

$\boldsymbol{d} = [t_0, t_1, ..., t_{|\boldsymbol{d}|-1}]$ Textual document composed of tokens.

$\mathcal{D}$ Domain. A collection (also called dataset or corpus) of documents $\boldsymbol{d}$ belonging to a specific domain. A domain is characterized by its type of text (e.g., encyclopedic, social network, news) and topic (e.g., biomedical, scientific, politics, music). As the topic influences the information presented in the documents, the domain also impacts entity types and relations types.

$\mathcal{E}$ Entity types. The set of entity types of $\mathcal{D}$.

$\mathcal{R}$ Relation types. The set of relation types of $\mathcal{D}$.

$\mathfrak{e} \in \mathcal{E}$ Entity type.

$\mathfrak{r} \in \mathcal{R}$ Relation type.

$\boldsymbol{m} = [t_{\text{start}(\boldsymbol{m})}, ..., t_{\text{end}(\boldsymbol{m})}]$ Entity mention located in $\boldsymbol{d}$.

$\boldsymbol{e} = (\mathfrak{e}, \{\boldsymbol{m}_0, \boldsymbol{m}_1, ...\})$ Entity of type $\mathfrak{e}$, composed of one or more mentions $\boldsymbol{m}$.

o Entity linking identifier. For instance, a Wikipedia URL, or a Wikidata ID.

$\boldsymbol{r} = (\mathbb{r}, \boldsymbol{e}_{head}, \boldsymbol{e}_{tail})$ Binary relation of type $\mathbb{r}$, linking $\boldsymbol{e}_{head}$ to $\boldsymbol{e}_{tail}$. Also called relation instance.

$\mathcal{G} = (\boldsymbol{E}, \boldsymbol{R})$ Knowledge graph $\mathcal{G}$ composed of a set of entities $(\boldsymbol{e}, o) \in \boldsymbol{E}$, and a set of relations $\boldsymbol{r} \in \boldsymbol{R}$.

$\mathcal{P}$ Prompt inputted to an EncLM, containing at least one `[MASK]` token.

I Introduction

Contents

The quantity of information created each year is exponentially growing. Estimated at 30 ZB (zettabyte)[1] in 2018, it reached 120 ZB in 2023, and is forecasted to attain 147 ZB in 2024 [12]. All that information is a goldmine for knowledge-centric and data-centric applications, but Gartner [13] estimates that only 20 % of this data is structured.

On the one hand, this 20 % is readily analyzable with specialized tools specifically designed to exploit this structure: 1. databases (SQL and NoSQL) for indexation, 2. that are easily searchable with specialized query languages built to benefit from this structure, and 3. visualization, analysis, and predictive tools covering the three descriptive, predictive, and prescriptive types of business analytics.

On the other hand, the remaining unstructured 80 %, composed of video, audio, and unstructured textual content, is vastly underexplored, with an estimated 0.5 % being processed and used [13]. In the context of this thesis, we restrict ourselves to textual documents. They contain valuable knowledge, but the absence of structure and the use of natural language render them particularly difficult to analyze. Indeed, natural language introduces ambiguity, as the same information can be expressed in multiple ways (formulation diversity), and a phrase can have different meanings depending on the context (polysemy or homography[2]):

[1] 1 ZB = 10^{21} B = 10^{9} TB. This quantity is so huge that it is difficult to fathom. To give an example, it represents 100 millions hours of 4K video content.

[2] A polysemic word is a *single* word that has multiple meanings depending on the context. In contrast, homographic words are *different* words (of different meanings) with the same spelling.

Formulation Diversity

- Bill Gates was born in Seattle.
- The birthplace of Bill Gates is Seattle.
- Bill Gates comes from Seattle.

All these sentences express similar information with different formulations.

Polysemy and Homography

- French persons speak French.
- Georgia the U.S. state, or Georgia the European country.

The first "French" refers to the nationality, and the second to the language. The case of Georgia is even more complex, as the two homographs are location concepts. Disambiguating them requires understanding the textual context to identify location hints.

This ambiguity makes unstructured documents challenging to analyze and integrate into an application as it requires prior language understanding and disambiguation steps.

Therefore, the scientific community deployed efforts toward specialized models to analyze unstructured documents to extract meaningful information. Once extracted, this information can be structured in databases, which are then easily integrated and processed with existing tools for structured data. This knowledge extraction and structuration task is called Information Extraction (IE). IE is the primary objective of this work: analyzing unstructured documents to extract and structure knowledge. However, IE is not new (in fact, it dates back to the first MUC[3] conference in 1987); many research directions have been explored, and it is currently exploding thanks to the rise of (large) language models [14, 15, 16]. The relevance of this thesis is linked to the industrial experimental constraints that originate from the practical needs of Alteca[4], which funds this work.

Indeed, Alteca is a French information technology services company specializing in digitalizing banking, insurance, and retail processes. In that context, it identified the customer's need to analyze various unstructured textual documents to extract and structure meaningful knowledge. This corporate and real-world context led us to consider strong constraints such as open-world extraction, minimizing the number of annotations required, document-level IE, and, above all, realistically usable and deployable models. We present and detail these constraints in the next section.

I.1 Context & Objectives

I.1.1 Unstructured Documents

We suppose to have access to a document collection (or dataset) $\mathcal{D}$ composed of multiple unstructured documents $\boldsymbol{d}$. Unstructured documents are strings with no layout or formatting

[3] Message Understanding Conference.
[4] Available at https://alteca.fr (in French).

metadata; in brief, plain text. We do not consider the task of extracting text from physical or digital documents and assume it has been done before our information extraction phase. Although we acknowledge the benefits of format and layout to improve extraction performance, the choice of entirely unformatted documents is motivated by two reasons.

First, the final objective is to work with company documents that may require digitalization or standardization steps. Notably, extracting text from PDFs or employing OCR on image documents is known to be ineffective or imprecise in preserving formatting and hierarchical information (titles, headers, emphasis, ...). Instead of relying on potentially missing, incomplete, and imprecise structural information, we choose to ignore this feature.

Additionally, Alteca has a partnership with Esker[5], a French company specializing in analyzing and extracting information from visually rich documents, such as invoices and bills. These visually rich documents follow a semi-structured layout (for instance, a tabular format), which they exploit to identify information precisely. Alteca favors their solution for semi-structured documents. As a result, this thesis aims to explore the analysis of lesser structured and more verbose documents, which are outside the scope of Esker's capabilities.

Finally, it is essential to notice that we work with document collections, that is, a large set of documents covering the same domain. It means documents that share similar topics, styles[6], or type of information. In practice, this corpus coherence is helpful for the extraction (see chapters IV and V).

I.1.2 Information Extraction

The base definition of information extraction is "extracting structured information from unstructured documents". This formulation is generalistic and covers a broad spectrum of methods and tasks: named entity recognition, relation extraction [17, 18, 19], coreference resolution [20, 21], event extraction [22, 23], sentiment or emotion analysis [24], or form filling [25, 26].

In the context of this work, we restrict ourselves to the task of Knowledge Graph (KG) construction, which aims at identifying the main concepts (also called entities) of documents and the relations that link them. Conceptually, it encompasses four tasks:

1. Named Entity Recognition (NER). It identifies the important concepts (or entities) of the document. The extracted entities are then grouped into entity types (*persons*, *locations*, *organizations*, etc.).

2. Coreference Resolution (CR). It merges the entities of a document that refer to the same concept.

3. Entity Linking (EL). It finds a unique ID for each entity (that can be seen as a database primary key). This step is necessary to generate structured and coherent knowledge that is integrable in existing databases.

[5] Available at https://www.esker.com.

[6] We do not imply similar structures, but similar language level or specialized vocabulary, that is, linguistical features specific to the corpus.

4. Relation Extraction (RE). It extracts the relations that are linking two entities in a document. The extracted relations are then grouped into relation types (*workplace of*, *born in*, *founder of*, etc.).

One noteworthy challenge linked to this formulation of IE is knowledge disambiguation. Indeed, coreference resolution and entity linking aim to solve the polysemy and formulation diversity particularities of natural language, to provide standardized and unambiguous extracted information readily integrable to existing (or new) databases. Without disambiguation, an IE model would most likely construct knowledge graphs and databases that contain duplicated nodes.

I.1.3 Industrial and Real-World Constraints

Document-Level Information Extraction Logically, as we speak of document collections, we want to extract information from documents. Surprisingly, most IE models are sentence-level and not document-level (see chapter II). At first glance, as a document is a list of sentences, one can hypothesize to process each sentence separately with sentence-level models to extract information. However, analyzing a large-scale document-level IE dataset, DocRED [1], demonstrates that 40 % of the relational facts are cross-sentence, meaning they require understanding multiple sentences jointly to extract them. Sentence-level models would miss such knowledge.

Handling complete documents instead of sentences introduces many questions and challenges, particularly 1. the complexity of encoding and keeping track of extended contexts, as well as 2. overcoming information scarcity (many entity or relation candidates but few valid ones).

Specific-Domain & Low-Resource The final objective for Alteca is to deploy our IE solution directly on the customer's premises, using its unstructured documents. The final business use cases, types of documents, and domains are not fixed a priori, and we aim for maximum flexibility. Therefore, we want IE models that can quickly adapt to the customer's domain.

This introduces challenges to adapt 1. to the specific target domain and 2. to the target language (most of Alteca's customers are French). Traditional IE models are supervised, requiring large amounts of documents from the target domain (and language) to be annotated. This process is costly and time-consuming, especially in demanding fields (law, economics, science, biomedical, etc.) where experts must be involved in annotation. Developing traditional supervised approaches raises the barrier to entry and may slow the adoption among Alteca's customers.

Therefore, this work focuses on low-resource approaches that minimize the labeling required to train the model. Multiple research directions can be studied: unsupervised approaches, zero-shot or few-shot models, active learning, self-supervision, synthetic data generation, etc.

Open-World Capabilities Given a domain-specific document collection, an expert user can define a structure (ontology) of information he is interested in extracting. This consists of specifying the list of entity types, relation types, and constraints (for instance, the relation

workplace of necessarily involves a person and a location). However, we believe being exhaustive in this scheme identification and structuration process is difficult, if not impossible. This incompleteness would result in missing meaningful and valuable knowledge.

Therefore, we want IE models that can extract and structure information automatically, even from entity types and relation types not initially envisioned by the user. In brief, open-world models. This constraint opposes the current tendency of models to be closed-world, for which entity and relation types must be specified beforehand.

Additionally, this open-world objective is linked to the low-resource constraint. Indeed, if entity and relation types are not exhaustively specified a priori, it is impossible to have annotated data covering all of them.

Real-World Use Finally, in this research work, we focus on methods that are applicable in practical scenarios. Even though it may seem a trivial and obvious constraint, we noticed, during our literature review, that some models or typologies of models were scientifically interesting and achieved impressive performances but were not usable in a real-world use case. The main reasons included:

- Computational complexity. Lee et al. [27] attain $\mathcal{O}(|\boldsymbol{d}|^4)$ complexity depending on the length of the document, making the extraction intractable with long documents.
- Execution cost. For each document and entity type, Xie et al. [28] require up to 80 large language model calls.
- Irrealistic experimental setup. Perez et al. [29] uncover that some few-shot models (supposed to be low-resource) need, in practice, large annotated validation datasets to adjust their hyperparameters.

The first two points are subjective and depend on the client's budget, but we have tried to propose models that have a good tradeoff between complexity/cost and performance. For the last point, we have been particularly cautious about experimental setups so that they are realistic and as close as possible to reality.

I.2 Research Directions, Main Contributions, and Publication Record

To summarize the context and constraints evoked previously, the main research question of this thesis is:

> *How to extract structured and open-world information from unstructured and domain-specific document collections in a realistic and low-resource experimental setup?*

This section briefly presents the main research directions we have followed to provide answers to this research question. Before starting, we remind the reader that Large Language Models

(LLMs) gained spectacular popularity from the end of 2022, that is, in the middle of our research work. This situation was equally a scientific opportunity and a significant obsolescence risk. Our earliest contribution (PromptORE) was published before the presentation of ChatGPT [30], and Linked-DocRED was built before the now widespread use of LLMs in IE [31].

I.2.1 Linked-DocRED: How to Evaluate Document-Level Information Extraction Models?

Our first step was to design a solid and objective experimental setup. This requires a complete set of metrics to evaluate the performances of IE models and a human-quality annotated dataset to train and evaluate methods. We came to the realization that such an experimental setup was missing. In particular, existing datasets were either incomplete (focusing on a subset of the tasks defined in section I.1.2), were too small, not diverse enough (few entity and relation types), or automatically annotated (without a strong guarantee of the correction of annotations).

Therefore, our first contribution was to propose Linked-DocRED, the first manually annotated, large-scale, document-level IE dataset. As creating and annotating a dataset from scratch was impossible due to time and budget constraints, we proposed enhancing the existing and widely-used DocRED dataset [1]. This dataset covers the NER, CR, and RE tasks but lacks entity linking labels. We generated them thanks to a semi-automatic process that guaranteed high-quality annotations. Indeed, realizing that DocRED documents were taken from Wikipedia articles, we designed an automatic algorithm to align DocRED instances with their corresponding Wikipedia page and mapped wikilinks with entities. Wikilinks are internal Wikipedia hyperlinks that redirect the user to the page describing the entity. As such, wikilinks can be considered unique IDs and, thus, valid entity linking labels. Moreover, as Wikipedia contributors manually edit wikilinks, they constitute a free yet human-quality source of entity linking labels.

Secondly, we also proposed a complete framework of metrics to benchmark end-to-end IE models, and we defined an entity-centric metric to evaluate entity linking by complementing the work of Zaporojets et al. [21]. These metrics were also extended to work under an unsupervised and open-world setting.

Finally, the evaluation of a supervised baseline IE model showed promising results while highlighting the numerous remaining challenges: 1. cascading errors inherited from the sequential nature of information extraction (e.g., relation extraction is based on the previously extracted entities), and 2. the complexity of handling long documents with distant entities and relations.

This first contribution led to the presentation of a resource paper at SIGIR'23:

- Pierre-Yves Genest, Pierre-Edouard Portier, Előd Egyed-Zsigmond, and Martino Lovisetto. "Linked-DocRED – Enhancing DocRED with Entity-Linking to Evaluate End-To-End Document-Level Information Extraction Pipelines". In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'23. Taipei, Taiwan: Association for Computing Machinery, 2023. ISBN: 978-1-4503-9408-6. DOI: 10.1145/3539618.3591912

I.2.2 PromptORE: Realistic Unsupervised & Open-World Relation Extraction

The previous Linked-DocRED contribution led us to two conclusions:

- A supervised and closed-world end-to-end IE model achieves F1 score performances under 60 %. This level of performance is not acceptable for a deployment use case.
- It is unlikely that low-resource and open-world models will attain better results.

Therefore, instead of pushing on an implausible successful end-to-end IE task, we decided to focus on specific IE subtasks to bring significant improvements. Our first choice turned to open-world and unsupervised relation extraction. This setting is particularly relevant for our constraints regarding specific domains where no annotated dataset is available and open-world extraction where relation types are a priori unknown. Although recent approaches achieved promising results, they heavily depended on hyperparameters whose tuning most often required labeled data, which is incompatible with a realistic unsupervised setting.

To diminish the reliance on hyperparameters, we proposed PromptORE, our Prompt-based Open Relation Extraction model. We adapted the prompt-tuning paradigm used in low-resource approaches to work in an unsupervised setting and used it to embed sentences expressing a relation. We then clustered these embeddings to discover candidate relation types and experimented with different strategies to estimate an adequate number of clusters automatically. To our knowledge, PromptORE is the first unsupervised and open-world relation extraction model that does not need hyperparameter tuning.

Experiments on one general and two domain-specific datasets showed that PromptORE widely surpassed previous state-of-the-art methods while being simpler and not needing any hyperparameter tuning. A qualitative analysis demonstrated that PromptORE identified most relation types without prior knowledge and provided a semantically coherent typing scheme.

This contribution led to the presentation of a long paper at CIKM'22, which was subsequently published at the national TALN'23 conference:

- Pierre-Yves Genest, Pierre-Edouard Portier, Előd Egyed-Zsigmond, and Laurent-Walter Goix. "PromptORE - A Novel Approach Towards Fully Unsupervised Relation Extraction". In: *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*. CIKM '22. Atlanta, USA: Association for Computing Machinery, Oct. 17, 2022. DOI: 10.1145/3511808.3557422. URL: https://hal.science/hal-03858264
- Pierre-Yves Genest, Pierre-Edouard Portier, Előd Egyed-Zsigmond, and Laurent-Walter Goix. "PromptORE – Vers l'Extraction de Relations non-supervisée". In: *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles*. 30e Conférence sur le Traitement Automatique des Langues Naturelles. Vol. 4. Paris, France, June 5, 2023, pp. 58–64. URL: https://coria-taln-2023.sciencesconf.org/459305

I.2.3 CITRUN: Extension to Open-World Named Entity Recognition

Encouraged by the successes of PromptORE, we then explored the second major task of IE, named entity recognition, under the same unsupervised and open-world setting.

In that NER task, we also observed weaknesses in the experimental setup of state-of-the-art unsupervised baselines. The majority of them were not open-world [35, 36, 37, 38], as they required supervision for each entity type. Conversely, zero-shot approaches that saw rapid progress were low-resource but assumed a closed world.

To tackle both shortcomings, we proposed CITRUN. We first adapted the unsupervised prompting method of PromptORE to recognize and classify entity types. We then complemented it by the use of off-domain labeled data, coming from generic domain datasets (CoNLL-2003 [39]) or synthetically annotated data (Pile-NER [40]). This off-domain data was employed to improve entity type embedding using contrastive learning. We experimentally observed that this embedding refinement step was beneficial, even when the domains were conceptually and stylistically very far away. For the entity detection subtask, we observed that the most straightforward token classification architecture was very effective, while more complex span-based models had lower cross-domain and generalization capabilities.

Experimental results showed that CITRUN significantly outperformed LLM-based unsupervised and open-world NER models while being 70 times smaller. Compared to the more supervised zero-shot NERs, CITRUN achieved competitive results.

The work of CITRUN is currently under review for publication.

I.3 Outline

Chapter II conceptualizes and formalizes information extraction and presents a literature review covering the recent advances and highlighting the main weaknesses and shortcomings at the core of our contributions.

The following three chapters focus on the contributions we have listed in the previous section. Chapter III introduces our Linked-DocRED dataset and its semi-automatic entity linking annotation, defines the set of metrics to evaluate IE models, and assesses a strong IE baseline. Chapter IV presents PromptORE, our prompt-based unsupervised and open-world relation extraction model. The core of PromptORE, its relation embedding module, is adapted for named entity recognition in chapter V, extended and complemented with cross-domain capabilities. This allows us to use off-domain or automatically generated data to train CITRUN, our NER model.

Lastly, we summarize our findings in chapter VI and highlight open questions and research areas for future work.

II Information Extraction

Technological Landmarks, Recent Advances, and Current Shortcomings

Contents

II.1 Introduction

Information Extraction (IE) aims to extract meaningful information from unstructured text and structure it to build or complement a Knowledge Graph (KG). The resulting knowledge graph

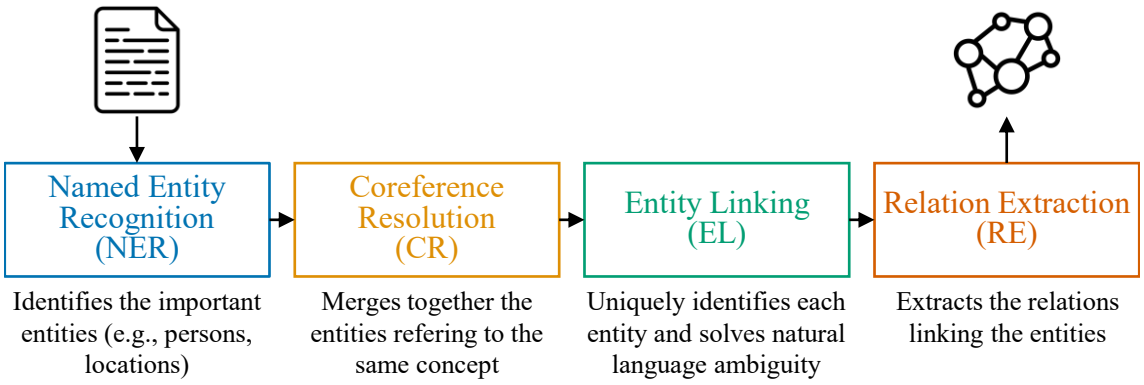


Figure II.1: Conceptual overview of information extraction.

can be used for multiple downstream tasks such as recommender systems [41], logical reasoning [42], or question answering [43]. As we have presented in the introduction, we restrict ourselves to the task of knowledge graph construction, which encompasses (see figure II.1):

1. Named Entity Recognition (NER). It identifies the important concepts of the document (persons, locations, organizations, etc.).
2. Coreference Resolution (CR). It merges the entities of a document that refer to the same concept.
3. Entity Linking (EL). It finds a unique ID for each entity (stable across documents).
4. Relation Extraction (RE). It extracts the relations that are linking two entities in a document.

These four steps can be done sequentially with an IE pipeline [44] or jointly with an integrated IE model [20, 45].

An analysis of more than 120 recent IE papers[1] shows that most of them consider only the NER and RE tasks, 20 of them consider CR (9 of them implement a partial[2] or implicit CR model), and only 8 consider EL [20, 23, 44, 46, 47, 48, 49, 50] (6 of them implement a partial[2] EL model). However, we believe coreference resolution and entity linking are important steps, as they transform ambiguous extracted triples into structured and disambiguated nodes and relations. The question of ambiguity in natural language is indeed essential. For instance, the surface form of an entity mention can refer to multiple entities: Georgia, the Eastern European country, or Georgia, the U.S. state. Conversely, an entity can be expressed with multiple surface forms: “Anakin Skywalker” and “Darth Vader”. CR and EL constitute the bridge between extracted triples that are ambiguous and structured knowledge that downstream applications can use. Ignoring these two steps (especially EL) hides an important part of the complexity of extracting information.

Writing a literature review about IE is a difficult endeavor, given the rapid pace of change in the domain and the multiplicity of challenges, such as handling long documents, working in specific domains without large annotated datasets, or extracting open-world information. This situation has led to a wide and diverse spectrum of approaches. To the best of our knowledge, recent surveys focus on a subset of IE, such as open-world IE [19, 51] (without the need to specify entity and relation types in advance), document-level IE [18] (in opposition to the more usual sentence-level IE), or LLM-based IE [52] (i.e., generative, in contrast with extractive IE). This chapter provides a general overview of IE without seeking to be exhaustive, focusing on the main methods, recent advances, and current shortcomings.

We start by defining and formalizing the conceptual tasks of IE (section II.2). We then present the major technological landmarks of IE (section II.3) and finish with particular focuses on document-level approaches (section II.4) and open-world methods (section II.5) that are

[1] We review these articles in the next sections.

[2] Methods based on large language models (e.g., [46, 47, 48]) benefit from their ability to handle coreferent pronouns and concept canonicalization that brings a certain kind of CR and EL. Still, it is implicit and unable to handle vastly different surface forms.

more applicable in a real-world scenario. This chapter highlights weaknesses and areas of improvement that constitute the core of our contributions.

Before starting, it is essential to re-contextualize our contributions to the field's history. In particular, the harness of LLMs capabilities to make predictions or to annotate and generate synthetic data, at the core of impressive advancements (e.g., universal IE, low-resource IE), are included in this literature review but were absent during most of our contributions.

II.2 Information Extraction Tasks

As we have said in the introduction, Information Extraction (IE) aims to analyze documents $\boldsymbol{d}$ from a dataset $\mathcal{D}$ to extract meaningful information (in the form of entities $\boldsymbol{e}$ and relations $\boldsymbol{r}$ linking the entities) to build a knowledge graph $\mathcal{G}$. $\boldsymbol{d}$ is composed of tokens t: $\boldsymbol{d} = [t_0, t_1, ..., t_{|\boldsymbol{d}|-1}]$. Tokens can be words, part of words, or punctuation, as defined by SentencePiece [11]. $\mathcal{G} = (\boldsymbol{E}, \boldsymbol{R})$ is composed of a set of nodes $\boldsymbol{E}$ and a set of edges $\boldsymbol{R}$.

Nodes of $\mathcal{G}$ Each node is a couple $(\boldsymbol{e}, o) \in \boldsymbol{E}$. $\boldsymbol{e} = (\mathbb{e}, \{\boldsymbol{m}_0, \boldsymbol{m}_1, ...\})$ is an entity composed of an entity type $\mathbb{e}$, and mentions (or surface forms) $\boldsymbol{m}$ that appear in one or multiple documents. A mention is defined by its text in the document: $\boldsymbol{m} = [t_{\text{start}(\boldsymbol{m})}, ...t_{\text{end}(\boldsymbol{m})}]$. o is a unique entity linking identifier (can be seen as a database table's primary key). It can be a number (e.g., a Wikidata identifier) or a string (e.g., a Wikipedia resource identifier). This unique identifier is usually missing in most IE models.

Edges of $\mathcal{G}$ Each edge is a relation $\boldsymbol{r} \in \boldsymbol{R}$. It is defined as a triplet $(\mathbb{r}, \boldsymbol{e}_{head}, \boldsymbol{e}_{tail})$ composed of a relation type $\mathbb{r}$, a head (or subject) entity $\boldsymbol{e}_{head}$, and a tail (or object) entity $\boldsymbol{e}_{tail}$.

IE can be subdivided into four conceptual tasks (see figure II.1) that we define in the following sections. As a side note, the distinction between these four tasks is conceptual. Prieur et al. [44] and Li et al. [46] consider independent modules that are pipelined together to form an IE pipeline, but many works have explored the possibility of combining these tasks together. For instance, many group together NER and RE [53, 54, 55, 56]. Kirstain et al. [57], Dobrovolskii [58], Lee et al. [59], and Zhu et al. [60] consider the joint NER and CR tasks. Joint CR and EL have been explored by Zaporojets et al. [61] and Agarwal et al. [62]. Finally, some IE models [20, 63] tackle NER, CR, and RE jointly. However, to the best of our knowledge, no integrated (not pipelined) model has been proposed to tackle NER, CR, EL, and RE jointly.

Additionally, the ordering of the tasks is arbitrary and can be modified in practice. For instance, entity linking can improve coreference resolution, or relation extraction can remove ambiguity during entity linking. Among others, Zaporojets et al. [61] propose integrating relational information extracted in $\boldsymbol{d}$ to improve coreference resolution and entity linking (thus positioning RE before CR and EL). Others [22, 64, 65] implement decoding algorithms to globally select the optimal set of relations and entities considering types constraints[3] (thus

[3]For instance, relations types impose entity types constraints: *workplace of* involves a person entity with a location entity. If a candidate triplet does not respect this constraint, it can be a hint that there is no relation or that one or both predicted entity type is not correct.

merging the predictions of multiple steps to increase the performance).

Named Entity Recognition (NER) It aims to identify the entity mentions $\boldsymbol{m}$ in $\boldsymbol{d}$. Most models (e.g., [39, 66]) also try to classify the entity type $\mathfrak{e}$ of each $\boldsymbol{m}$. The list of entity types $\mathcal{E}$ can be known in advance or determined by the model (open-world NER).

Coreference Resolution (CR) In a document, multiple entity mentions can refer to the same underlying entity; they are called coreferences. Coreference resolution aims to merge coreferent mentions in a single entity $\boldsymbol{e} = (\mathfrak{e}, \{\boldsymbol{m}_0, \boldsymbol{m}_1, ...\})$.

In practice, Zeldes et al. [67] and Porada et al. [68] observe that there are multiple forms of coreferences:

- Exact match. Two coreferring mentions that have exactly the same text. At first glance, this case may seem trivial. However, multiple mentions can share the same text, yet not corefer:

  "A *French* person speaks *French*."

  Here, the first French refers to the nationality, and the second to the language.

- Modifiers. The coreferring mention is a modifier, that is, a word or a group of words that modify the meaning of a noun (similarly to an adjective). For instance:

  "*Taiwan* authorities have decided..."

  Here, *Taiwan* is a noun, but its function is that of an adjective.

- Generics. The coreferring mention is a generic noun phrase. For instance:

  "Microsoft released its quarterly results yesterday. *The company* has achieved its yearly revenue goals."

- Pronouns. The coreferring mention is a pronoun referring to an antecedent mention.

  "Microsoft released its quarterly results yesterday. *It* has achieved its yearly revenue goals."

This variety of coreferences types has an impact on evaluation. For example, OntoNotes [69] (one of the most widely used coreference datasets) considers all the previous coreference types, but PreCo [70] or DocRED [1] do not annotate pronouns.

Another challenge of CR is the complexity of handling long documents because the longer the document, the more entity mentions (that can corefer) it contains; and the number of candidate coreference relations increases quadratically. This necessitates efficient processing such as coarse-to-fine reasoning [27, 57, 58, 59], which aims to quickly and cheaply prune most coreference candidates to focus on a few promising ones, for which more complex models are used to score.

Entity Linking (EL) Entity linking can be seen as the generalization of coreference resolution in an inter-document setting[4]. Indeed, the objective of EL is to group entities $\boldsymbol{e}$ referring to the same concept but occurring in multiple documents. Conversely, entity mentions that share the same text but are not the same concept should not be merged. The objective is generally to find a unique identifier o describing each entity $\boldsymbol{e}$, such as a knowledge graph ID.

A typical EL model consists of three stages:

1. Candidate generation. It finds the identifier candidates for a given entity $\boldsymbol{e}$. The methods are based on information retrieval techniques related to search engines [71].

2. Candidate ranking. It ranks the candidates from the most probable one to the least probable. It usually involves heuristics (such as the popularity of the candidate) [72, 73], natural language processing treatments [61], or knowledge graph features [73, 74, 75].

3. Non-existing candidate prediction. It determines if the first KG candidate corresponds to the entity or if a new node has to be created.

Relation Extraction (RE) Relation extraction aims to identify the entities $\boldsymbol{e}_{head}$ and $\boldsymbol{e}_{tail}$ that are linked by a binary relation[5] $\boldsymbol{r}$ and find its type $\mathbb{r}$. Similarly to NER, the list of relation types $\mathcal{R}$ can be known in advance or automatically determined by the model (open-world RE).

II.3 Overview of Information Extraction

II.3.1 Pre-neural Information Extraction

Information extraction is rooted in the context of the seven Message Understanding Evaluation and Conferences (MUC) between 1987 and 1997 [77, 78, 79, 80, 81, 82]. They were financed by the U.S. Defense Advanced Research Projects Agency (DARPA) and aimed at automatically analyzing military messages, such as fleet reports, airplane crash reports, or rocket/missile launches, to extract specific information.

Early approaches are "triplet-based". They aim to extract triplets of the form (head, predicate, tail), with *head* and *tail* the surface form of the head and tail entities, and *predicate* the surface form of the relation. These triplets can then be decomposed into entities and relations. The main weakness of these triplet-based approaches is the complexity of canonicalizing predicates [83]. The same conclusion can be said about entities: these models do not implement coreference resolution or entity linking mechanisms, leading to duplicate nodes.

Most early methods rely on grammatical features: they often analyze pos-tagging features and syntactic or dependency trees of documents to identify relation triplets. Triplets are extracted using handcrafted rules [53, 84] or automatic algorithms [54]. With TextRunner, Banko et al. [54] are the first to speak of OpenIE, an IE setting where entity and relation types are not predefined. More recently, improvements have been made to simplify the documents to improve

[4] As a matter of fact, some approaches tackle entity linking and coreference resolution jointly [61, 62].

[5] Some works (such as [76]) study n-ary relations, but most approaches restrict themselves to binary relations.

extraction [83, 85, 86, 87, 88], for example, by splitting multi-phrase sentences, separating conjunctive clauses, or paraphrasing the document.

II.3.2 Neural Information Extraction

With the emergence of word embeddings and language models [14, 89, 90], purely grammatical approaches were mostly dropped in favor of deep-learning models, with two leading architectures: CNNs and RNNs (especially LSTM [9] and GRU [91]). These two architectures have proven successful thanks to their ability to integrate textual context information automatically [92].

Zeng et al. [93] and Liu et al. [94] propose using CNNs to extract relations. The major difference with previous approaches is that they do not rely on manually engineered features but on word embeddings and achieve better results. CNNs were later complemented with attention mechanisms [95] or external knowledge information [96].

Clancy et al. [49], Dligach et al. [92], Li et al. [97], and Manning et al. [98] implement LSTM-based networks to extract relations and achieve similar or better results than CNN-based methods. Recurrent models were also complemented with attention mechanisms [97, 99]. In the connex domain of open-world IE, Stanovsky et al. [55] propose RnnOIE, the first deep learning model that outperforms rule-based open IE models. This approach is further refined by SpanOIE [100]. A notable paradigm that appeared with LSTMs is graph-based IE [20, 64, 76, 101, 102]. They propose to model words of a document and corresponding entity and relation candidates into a graph. The embeddings of the nodes and edges are extracted from LSTMs [101, 102] or calculated by applying principles similar to message passing of graph neural networks [8, 76]. Reasoning algorithms then use this graph to identify relations. This formulation makes it possible to encode long documents and extract distant relations. Going one step further, Verlinden et al. [20] propose injecting existing knowledge from a KG into their IE model (using a soft entity linking module) to improve the extractions. KG embeddings are computed using GloVe embeddings of nodes' descriptions or graph embeddings (with TransE [75]). Finally, DyGIEE [64] proposes jointly learning NER, CR, and RE using a Bi-LSTM network with multi-task learning. They empirically observe that mutualizing knowledge across tasks improves performance and generalization compared to previous pipelined approaches.

Contrary to rule-based models that generate triples, most CNN and LSTM models are classifiers[6], meaning they tag or classify document tokens or spans to output predictions (entities or relations). A typical classification paradigm for NER is BIO. Each token is classified as *O* (not an entity), *B* (first word of an entity), or *I* (second or following word of an entity). *B* and *I* can be specialized for each entity type (e.g., *B-location*, *I-person*). In general, relation extraction or coreference resolution models are classifiers that take in input two previously extracted entities [92, 94]. The problem with this formulation is that every pair of entities has to be tested, which is computationally expensive. As a result, Stanovsky et al. [55] and Yuan et al. [99] adapt the BIO tagging scheme for relations, leading to faster inference (often at a cost in quality). Similarly, Lee et al. [27, 59] and Yu et al. [103] propose to prune most relation/coreference candidates using a fast model and then refine the most probable ones using more resources in a secondary step.

[6]Except Cui et al. [56] that we discuss in section II.3.4.

CNN models are computationally efficient and good at representing local context [92], but they struggle to encode long documents. RNNs can encode longer contexts but are inefficient due to their sequential nature [15]. The emergence of transformers [15], solely based on the principle of attention, and notably Encoder-Only Language Models (such as BERT [6]), opened a wide range of possibilities.

II.3.3 Encoder-Based Information Extraction

Previous architectures and paradigms have been reused and refined using Encoder-Only Language Models (EncLM). In particular, DyGIEE++ [22] builds upon the success of DyGIEE [64] by replacing Bi-LSTM with BERT and formally defining and refining the entity, coreference, and relation graph using graph neural networks (GNN) [8]. OneIE [104] prunes the graph to respect type compatibility constraints (e.g., removing relations between incompatible entities). Nguyen et al. [105] improve OneIE by using graph convolutional networks [7]. EnriCo [65] adopts an alternative soft graph decoding method that globally finds the optimal set of entities and relations that respect the type compatibility constraints. Finally, Zaratiana et al. [106] replace the usual message passing algorithm to update GNN embeddings by TokenGT [107] (representing nodes and edges using transformers encoder layers), allowing for more expressive embeddings and state-of-the-art results.

Tagging [40, 108, 109, 110] and span-based [111, 112, 113] classification formats have also been adapted for BERT. Similarly to RNN-based models, efforts have been made to decrease the computational complexity of span-based approaches. One notable idea is table-filling: Wang et al. [114], Ma et al. [115], and Wang et al. [116] employ bi-affine layers to fill a 2-D table (with tokens of the input sentence in the axes). A high score on a specific cell (i, j) indicates a connection between token t_i and t_j. For NER, it indicates an entity spanning tokens i to j. For RE or CR, it signifies a relation between an entity starting at t_i and another entity starting at t_j. TablERT-CNN [117] further shows that it is possible to maintain high performances while freezing the EncLM parameters and using a shallow CNN. Finally, Yan et al. [118] proposes the "PlusFormer" to replace the usual bi-affine layers. The PlusFormer generalizes the self-attention mechanism in a table, considering the interactions between token pairs.

A novel type of IE model is based on machine reading comprehension (MRC) [119]. MRC models are designed to answer questions about a text. Li et al. [36] propose to ask an EncLM question-answering model to extract and type entities. Zhao et al. [108] obtain promising results for relation extraction by asking diverse questions (paraphrasing) and aggregating the answers with a weighted voting system.

Most of the proposed IE approaches are joint and trained thanks to multi-task learning, with the idea that sharing knowledge between tasks is beneficial [22, 104, 108, 111]. However, PURE [113] shows empirically that two simple span-based NER and RE models, trained separately and pipelined together, obtain state-of-the-art results while being conceptually easier and faster to train.

Another area of interest is pre-training (or adapting) language models specifically for IE tasks. A foundational work is ERNIE [120], which complements the usual masked language modeling (MLM) pre-training task with an entity modeling task. In this task, entire entities (composed

of one or more tokens) are masked to improve the contextual association between entities and their context. Indeed, it is very easy to predict *San* in the sentence "The Golden Gate Bridge is located in [MASK] Francisco."; but it requires more contextual knowledge integration to predict *San Francisco* given "The Golden Gate Bridge is located in [MASK] [MASK]." Similarly, Soares et al. [121] propose extending this task to pairs of entities to improve relation embeddings and Wang et al. [112] to two other IE tasks (relation typing and entity typing). These models achieve better results than raw BERT embeddings for IE tasks and can be used as drop-in replacements. With the recent advances of LLMs and their impressive generative capabilities, Bogdanov et al. [122] and Peng et al. [123] propose to use GPT [124] to annotate data and use these silver labels to fine-tune BERT embeddings and provide EncLMs that can be used as backbones for IE tasks.

LLMs are also employed by Naraki et al. [125] to complement the annotations of manually labeled data (to improve recall) or by Ye et al. [110] to augment few-shot exemplars by paraphrasing. Ma et al. [109] propose using LLMs to refine complex entity and relation type predictions. The most probable classes are presented to an LLM as a multiple-choice question, which selects the correct answer.

Finally, Lou et al. [40] and Zaratiana et al. [66] propose "universal" or general-purpose IE models that are trained on a wide variety of manually labeled [40] or automatically generated [31, 66] datasets. The schema (list of entity types and/or relation types) is specified in the input, and they are often used in a zero-shot setting, making them ideal when annotated data is scarce or non-existent. To clarify, universal IE is rooted in generative IE, particularly LLM-based IE, which we review in section II.3.4. However, these papers demonstrate that strong generalizability and performance are attainable with small EncLMs.

II.3.4 Generative Information Extraction

Previously presented models are extractive: they tag, classify, or identify spans inside a document to extract information. In contrast, generative IE models generate text, code, or augmented languages, which are then parsed to output predictions. This prediction format has recently gained popularity due to the rise of LLMs and the impressive results they obtain on several NLP tasks [28, 63, 126]. However, the foundational contributions toward generative IE predate LLMs (and even BERT). Indeed, in 2018, Cui et al. [56] proposed a Bi-LSTM IE model based on the seq2seq architecture [91, 127] that generated the extracted information (relation triplets) as a natural language output. This model was perfected by IMoJIE [128], which replaced the Bi-LSTM encoder with BERT (while keeping an LSTM decoder). They also mitigate the "stuttering problem" of [56], which tends to repeat identical triplets or generate very close ones, by conditioning the following triple on all previously extracted. REBEL [129] employs virtual tokens to better separate head, predicate, and tail arguments. GenIE [130] investigates beam search to force the model to follow a predefined output scheme and reduce hallucinations. Additionally, TANL [131] implements the Needleman-Wunsch algorithm [132] to align the generated triples with the input text and correct them, if necessary, to reduce the impact of hallucinations.

Built upon the successes of these foundational models, the domain of generative IE is in full expansion, and there is no clear consensus on the most efficient design for an IE model. We list

dichotomies that are currently explored:

- Full transformers (T5 [133, 134], BART [135]) vs. decoder-only (GPT [124, 136, 137, 138], Llama [16]). Decoder-only LMs have better scalability than full transformers, but recent papers [139] show that small transformers can outperform larger decoders trained on the same data.
- Frozen vs. fine-tuned. Smaller fine-tuned LLMs are ideal to run locally and attain impressive IE performances [23], but clever prompting of frozen large LLMs can achieve similar performances [28].
- Natural language vs. code. Code-LLMs [140] and LLM-based IE models are neck and neck in terms of performance [23, 50, 141]. As a side note, foundational models relied on "augmented languages" that define additional virtual tokens to separate arguments or model classes. This paradigm has been abandoned (except for ATG [63]), and recent approaches use, on the contrary, formats that are as close as possible to the languages seen during pre-training.

Instead of employing augmented languages, Townsend et al. [26] and Lu et al. [47] propose to follow a JSON format to benefit from an already known syntax. By doing so, they improve performance compared to augmented language models. Natural language is also used by UIE [24], which is the first universal IE model[7]. UIE combines two essential ideas: 1. specifying the task and the list of classes in input (using a structure schema instructor), and 2. training the model on various datasets covering diverse domains. UIE can then be applied to new domains without annotated data, following a zero-shot setting. Fei et al. [142] improve UIE by merging grammatical features (syntactic and dependency trees) in the embeddings. A weakness of UIE underlined by Wang et al. [143] is that it produces task-specific models (for NER or RE), not a general IE model. As a result, they propose InstructUIE, which is fine-tuned once for a wide variety of IE tasks, and they experimentally observe that this joint training is beneficial for task-specific performances. Ling et al. [144], Zhang et al. [145], and Wang et al. [146] propose to implement a secondary answer verification and error correction step by prompting an LLM a multiple choice question (focusing on challenging classes) and asking for explanations (Chain of Thought [147]). Finally, negative sampling during training is an essential factor in the performance of universal IE models [31, 50, 139]. Negative sampling randomly chooses entity (resp. relation) types absent in the currently considered text to ensure the model answers an empty list. Notably, Zhou et al. [31] find that frequency-based negative sampling (taking into account the frequency of occurrence of the type) is more effective than uniform negative sampling.

The use of Code-LLMs for IE was initiated by CodeIE [46]. They make the hypothesis that programming languages naturally implement 1. a structured format that is easily parsable and 2. structural concepts ideal for specifying a schema (e.g., classes, comments, inheritance, typing). Experimentally, they perform better than the natural-language-based UIE [24]. This idea is perfected by Code4UIE [148] and Code4Struct [149] that implement in-context-learning

[7] It predates USM [40] we have mentioned in section II.3.3.

to select few-shot exemplars and propose a dual-stage model that first identifies the mentioned types (using the Python import syntax) and then predicts entities and relations associated to the previously identified classes. Bi et al. [150] ask the Code-LLM to generate extraction explanations in code comments. GoLLIE [23] and KnowCoder [48] further enhance CodeIE by adding the concepts of inheritance (to model hierarchy of types), type hints (to model type constraints for head and tail entities of a relation), or class comments (for type descriptions and zero-shot examples).

Frozen LLMs are not out of the picture. A method that stands out from the crowd is ChatIE [151]. They propose a dual-stage raw prompting method that first probes the list of entity (resp. relation) types expressed in the document and then iteratively identifies entity (resp. relation) belonging to each type[8]. Sun et al. [141] propose automatically annotating in-context-learning exemplars with GPT and using them in the prompt. Although the annotation of exemplars is imperfect/automatic, they observe that it improves performance compared to prompting without them. In addition, Xie et al. [28] implement an ensemble method (with different prompts and sampling parameters) to refine exemplars annotations, resulting in scores equivalent to those of fine-tuned approaches[9].

A recurring concept with generative IE is Chain of Thought [147]. The model is asked to explain each extracted information (entity, relation, or triplet). This process positively impacts frozen or pre-trained LLMs [126, 152, 153, 154]. Interestingly, Wadhwa et al. [152] propose to generate an explanation by a large LLM (e.g., GPT-3 [155]) and fine-tune a small transformer (T5) to reproduce this explanation, leading to improved results.

An area where generative IE particularly shines is low-resource IE. We can summarize the contributions in three directions.

First, universal IE (e.g., [23, 24, 48, 143], see previous paragraphs) proposes general-purpose IE models that are trained on a wide variety of manual or synthetic datasets and for which the user can specify the extraction scheme (classes and structure). They can be applied to new domains, with previously unseen types, and without annotated data (zero-shot) while still performing honorably.

Second, LLMs allow the creation of synthetic data of unseen size and quality (compared to distant-supervision [156, 157]). This data can be used as in-context-learning exemplars [50, 141], or to fine-tune IE models (e.g., UniversalNER [31], NuNER [122]). It is interesting to notice that naive synthetic document generation does not work. Indeed, Josifoski et al. [158] observe that, without constraints, it is impossible to obtain diverse (in terms of domains, entities/relations, types) documents valid for training. To tackle this problem, Zhou et al. [31] and Bogdanov et al. [122] propose to use actual documents (taken from the Pile Corpus [159]) and annotate them with an LLM, ensuring domain coverage. Josifoski et al. [158] take the opposite direction: they generate realistic entity/relation graphs and prompt an LLM to create the text corresponding to the graph.

Finally, with their low-resource learning and generation capabilities, we see recent human-

[8] As an aside, it is interesting to notice that they reverse the usual extract then type process employed by most non-generative approaches.

[9] One weakness of their approach is that it is costly as it requires numerous prompting per sentence.

in-the-loop applications. For instance, InteractiveIE [160] structures information with LLM prompting in an unsupervised setting and involves a human annotator to merge or split clusters at a high level that are then used for subsequent extractions. Dagdelen et al. [25] propose manually annotating a small sample of data. This small amount of data is used to bootstrap an active learning[10] process where an LLM is fine-tuned on documents and automatically annotates new documents checked and complemented by a human annotator. This method seems promising in specific domains where universal IE struggles.

Is Information Extraction Solved by Large Language Models? [161] With everything we have mentioned in the previous paragraphs, we can only observe that LLMs have invaded IE. Although the proposed methods are conceptually and technically simple, they achieve impressive results, particularly for low-resource and universal IE. Hence, we raise the question, as do Han et al. [161]: do LLMs solve information extraction?

It is impossible to answer definitively, as one cannot foresee the future, but we want to highlight two discussion points.

First, regarding raw performance against other architectures, Han et al. [161] observe that generative IE outperforms encoder-based models on simple tasks (NER, sentence-level RE) but fails on more complex ones (fine-grained NER, document-level RE). In particular, for RE, most generative IE models [23, 153] are evaluated on ACE04[11], ACE05[12], SciERC [64] or CoNLL-2004 [162]. These datasets have few relation types (6 for ACE04 and ACE05, 7 for SciERC, or 5 for CoNLL-2004) and contain only sentences. The tests done on complex RE datasets, for instance, DocRED [1], which contains documents annotated with 96 relation types, show that generative IE falls behind encoder-based IE. The question of size is also a critical factor: the small performance gains for easy tasks brought by LLMs do not counterbalance their large size for every user.

However, we believe hybrid approaches combining generative and conventional IE may be a more reasonable short-term future. For example, Zaratiana et al. [63, 66] and Ding et al. [139] show that small encoders or transformers attain state-of-the-art-results when trained on LLM-annotated data. Independently, Ma et al. [109] find that LLMs can help to refine predictions on arduous instances. They advocate the use of small models to decide most of the cases and LLMs only when there is uncertainty or ambiguity.

The second discussion point may be more fundamental. Information extraction historically exists because of the need to *explicitely* extract and structure information that downstream applications can easily use. It is, in fact, the primary motivation of this thesis. However, with the development of LLMs that 1. encode a large amount of general [163] or specific [164] knowledge, and 2. integrate contextual knowledge in their answer thanks to, for instance, Retrieval Augmented Generation (RAG) [165], one can imagine a fully *implicit* knowledge pipeline. In that case, knowledge is never explicitly extracted nor structured into a KG but implicitly taken into account by the LLM when answering a question. Similarly to Pan et al.

[10] Active learning is a bit misleading. No particular instance selection strategy was implemented other than random sampling. But the general active learning select, predict, correct, and improve procedure is implemented.

[11] Available at https://catalog.ldc.upenn.edu/LDC2005T09.

[12] Available at https://catalog.ldc.upenn.edu/LDC2006T06.

[166], we believe such an implicit pipeline is currently unreasonable, in particular, due to the hallucination problem [167, 168]. On the contrary, KG seems to be beneficial for LLMs, as incorporating structured knowledge during the RAG process improves the overall result [169, 170, 171].

II.4 Document-Level Information Extraction

Document-level IE introduces two major challenges:

- Long documents. Encoding and understanding a long context is a difficult task. Additionally, relations can span multiple sentences, and therefore, conventional sentence-level IE may miss those relations. Indeed, an analysis of the document-level DocRED [1] dataset shows that more than 40 % of relations are multi-sentence and cannot be captured by sentence-level models.

- Combinatorial explosion. The number of relation candidates evolves quadratically depending on the number of entities. Logically, a document of n sentences contains n^2 times more relations candidates than a single sentence. Most candidates are not true relations: a document-level IE model is confronted with high scarcity [172].

In practice, due to the complexity of the task, most works focus only on relation extraction and assume that entities and coreferences have been perfectly extracted (except for [126, 141, 173, 174, 175] that we discuss in the last paragraph). To our knowledge, the first document-level RE model was created by Swampillai et al. [172]. It was based on the grammatical analysis of the sentence with an SVM classifier. They highlighted the problem of high levels of negative candidates and proposed learning a threshold to adjust the predictions accordingly. Currently, methods to handle long context and scarcity can be divided into graph-based and sequence-based approaches.

Graph-based RE DISCREX [101] proposes organizing the document in a graph, where nodes are words of the documents, edges are syntactic dependencies between words, links between words and their sentences, and coreferences. Relations are predicted by reasoning over the graph. Peng et al. [76] propose to refine node and edge representations using graph LSTMs. EoG [102] changes the perspective of the graph; the nodes are now entities/mentions, and edges are coreferences, mention to entity links, or mention to sentence relations. Zeng et al. [176] introduce dual graphs, with a mention-level graph (similar to [102]) and another higher-level entity-level graph. This allows them to represent low-level information and to structure and refine it in the higher-level graph (e.g., entity embeddings are the average representation of their mentions). Embeddings are updated thanks to graph convolutional networks [7]. Zhang et al. [177] explore multi-hop reasoning to tackle relations where entities are never mentioned in the same sentence. Finally, POR [174] pushes this multi-hop reasoning further with attention mechanisms and breadth-first search.

Sequence-based RE Graph-based approaches are complex and slow at inference. As a result, Zhou et al. [178] (ATLOP) explore the possibility of attaining state-of-the-art results without using graphs. To mitigate the scarcity problem, they propose to compute a per-instance prediction threshold and implement an attention mechanism to model the interaction between subject and object entities. In another direction, PAEE [179] proposes a coarse-to-fine approach to quickly filter most entity pairs that are not related (using pair-aware representations) to focus on the most probable relation candidates.

Huang et al. [180] are the first to propose evidence-based RE. They postulate that, given a relation candidate, reducing the context to focus only on the essential sentences for this candidate improves the results. To do so, they propose simple heuristics to select at most three sentences for a given candidate. This method is perfected by Xie et al. [181] and Ma et al. [182], who model evidence extraction jointly with RE and learn it multi-task (going further than simple heuristics). Xu et al. [173] and Xiao et al. [183] propose jointly training NER and CR with evidence extraction and RE, anticipating that RE can benefit from these subtasks. Finally, DocGNRE [184] enriches document-level datasets with automatically generated documents (with GPT-3), as well as filtering extracted relations with a natural language inference model [185].

Finally, Verlinden et al. [20] and Wang et al. [186] explore the injection of external knowledge about entities (coming from a KG) using entity linking. In particular, KIRE [186] is directly pluggable on existing RE models (such as ATLOP [178]) and consistently brings performance improvements.

Recently, some methods explored the joint document-level NER and RE tasks. Dong et al. [187] apply a similar principle as the generative IE model of Cui et al. [56]. JEREX [188] proposes multi-task learning for NER, CR, and RE. Zhang et al. [145] unifies the CR and RE tasks together (as coreference can be considered a specific type of relation). They employ a table-filling approach [115] to reduce the computational complexity and represent candidates in a graph whose embeddings are refined using graph neural networks. Finally, Xue et al. [126] and Sun et al. [141] explore the applicability of LLMs for joint NER and RE. GenRDK [141] implements in-context-learning with GPT-generated examples (and thus is working in a complete zero-shot setting). Xu et al. [52] test different ways to formulate the task and fine-tune small LLMs. They observe that the best approach is first to identify relation types, then the subject entities, and finally, object entities (three prompts). However, a performance analysis shows that these methods struggle to surpass $50\,\% - 60\,\%$ in F1 score, which makes them unusable in real-world scenarios. This demonstrates the challenges linked to document-level IE.

Finally, to the best of our knowledge, end-to-end document-level IE, covering the four NER, CR, EL, and RE tasks, has not been explored. We believe the major factor that hindered such a study is the lack of high-quality training and test datasets. This is the main motivation of our first contribution, Linked-DocRED: creating the first document-level, large-scale, human-quality dataset and implementing and evaluating an end-to-end IE pipeline.

II.5 Open-World Information Extraction (OpenIE)

Conventional IE models make a closed-world hypothesis: entity and relation types must be specified beforehand. It is logical for supervised approaches, as they need annotated data for each class, but even the very low resource zero-shot models need to know the type schema. This property is problematic for KG construction. Indeed, as evoked in the introduction, it is relatively easy to define a reasonable entity and relation type schema for a KG. However, it will not be exhaustive, meaning some documents may contain knowledge that the user is unaware of and have not been structured in its proposed scheme. With conventional closed-world IE models, such knowledge will not be extracted. This is the primary motivation for open-world IE. In this setting, the scheme is not predefined by the user but identified by the model. The general objective is to build a "universal" IE model that extracts all possible knowledge and proposes a reasonable typing scheme to structure it in a KG. Two families of models represent OpenIE: triplet-based and clustering-based approaches.

Triplet-based OpenIE Triplet-based methods aim to extract (head, predicate, tail) triplets from the document. Banko et al. [54] were the first to define the term OpenIE (although prior works by Etzioni et al. [53] and Shinyama et al. [84] were already open-world). Early OpenIE was based on grammatical features. KnowItAll [53] implements predefined, generic, grammatical rules for open-world IE, and TextRunner [54] proposes a self-learning framework where the model automatically finds new rules to extract more knowledge. These methods were refined with more deep grammatical features: dependency tree [85], verb constraints [86], phrase splitting [87], conjunctive sentence simplification [88]. Finally, SenseOIE [189] proposes to combine previous OpenIE in an ensemble method.

DocOIE [187] and IMoJIE [128] introduce generative OpenIE. While grammatical OpenIE extracts span from documents to create triplets, they propose to generate the triplet in an auto-regressive fashion. Their work is a generalization of Cui et al. [56] to unknown entity and relation types. IMoJIE is trained on silver annotations generated with the previous state-of-the-art OpenIE models (this idea has been reused since by Kolluru et al. [190] and Nagumothu et al. [191]). More recently, LLMs have been applied to generative OpenIE. Compared to generative closed-world IE, they are triplet oriented and do not require specifying types [47, 144]. Ling et al. [144] introduce an error correction mechanism with a dual extract then refine process. PIVOINE [47] proposes using large-scale, automatically annotated data to pre-train an LLM specifically for open-world IE tasks.

Generative OpenIE is powerful and conceptually elegant but costly to run. Indeed, triplets are sequentially generated one after another. Therefore, Stanovsky et al. [55] and Kolluru et al. [190] propose using word embeddings and sequence tagging to identify triplets. Their method (OpenIE6) is much quicker than IMoJIE [128] while maintaining a similar level of performance. OpenIE6 is enhanced by PIE-QG [191], thanks to sentence simplification. They paraphrase the input sentences using PEGASUS [192] and replace pronouns with their coreferent names. To further fasten the extraction process, MacroIE [193] adopts a coarse-to-fine approach with a primary module that quickly prunes most of the head and tail candidates and a secondary module that refines the remaining candidates to output the predictions. To do that, they represent

potential candidates in a graph and find triplets of tightly linked nodes (maximal cliques identification). Ro et al. [194] separates entity and predicate identification and learns them in a multi-task and multilingual setting. Finally, Dong et al. [195] (SMiLe-OIE) propose considering syntactic features and fusing them with BERT embeddings thanks to multi-view learning.

The main weakness of triplet-based OpenIE is the lack of predicate disambiguation. Indeed, as they rely on surface forms for the predicates, they likely produce duplicated predicates for the same underlying relation (e.g., "born in" and "birthplace of"[13]) that cannot be reconciled using grammatical rules.

Clustering-based OpenIE To solve the triplet-based OpenIE shortcoming, clustering OpenIE was proposed. Instead of relying on surface forms, clustering models generate clusters of close relations (or entities) that are expected to have the same type. This category of models is studied in more detail in section IV.2. We recall only the main elements in this paragraph. The methods usually follow a dual-stage architecture with:

1. Entity/Relation type embedding. The objective is to compute, for each candidate, a vector representative of the entity or relation type.

2. Type clustering. It aims to group candidates that share the same relation or entity type.

Language models are at the core of these approaches, providing expressive and precise relation or entity-type embeddings for each candidate. The whole clustering or grouping part is useless if this embedding is incorrect. Wu et al. [196] propose to refine relation embeddings generated with EncLM (the concatenation of the embeddings of the head and tail entities) thanks to contrastive learning on supervised data. To not rely on supervised data, SelfORE [197] implements an iterative self-supervision approach: it generates pseudo-labels (from clustering) that improve EncLM representations, which generates new embeddings that are then clustered to produce new pseudo-labels. RoCORE [198] extends SelfORE when labeled data is available (using semi-supervised learning). Duan et al. [199] observe that replacing SelfORE hard clustering with soft-clustering to generate soft pseudo-labels benefits the performances. Web-Selfore [200] proposes to gather information about the head and tail entities from the web (KG or Wikipedia) to enhance the relation embedding. Wang et al. [201] implement EncLM prompting to generate the relation embeddings[14] instead of the entity pair representations (e.g., [196, 197]), with a procedure similar to RoCORE and obtain state-of-the-art results. Recently, ASCORE [202] generalizes the active learning process for OpenIE. The user is asked to annotate selected samples manually, chosen to be in areas of maximum density and to cover the whole latent space.

These clustering-based OpenIE methods are more applicable in real-world use cases than triplet-based ones because they can group predicates with very different grammatical expressions. They are at the origin of two of our contributions, PromptORE for RE (chapter IV) and CITRUN for NER (chapter V).

[13] This case is even more complicated than a simple surface form difference, the entities are inverted: (Bill Gates, born in, Sealtle) or (Seattle, birthplace of, Bill Gates).

[14] With PromptORE, we discovered a similar prompting encoding approach. The works were done concurrently: PromptORE was published in CIKM'22, and MatchPrompt at EMNLP'22 two months later.

II.6 Conclusion

To conclude, information extraction is a central NLP task that aims to extract meaningful knowledge (composed of entities and relations linking these entities) from unstructured documents to build a knowledge graph. This knowledge graph is more accessible for downstream applications to use than initial documents that are ambiguous. IE is composed of four conceptual subtasks: named entity recognition, coreference resolution, entity linking, and relation extraction. A synthetic literature review shows that this topic is in the midst of a revolution thanks to language models (encoder-based and generative), and promising results are obtained, particularly in low-resource settings that were once very difficult. However, multiple challenges remain:

- End-to-end IE. The end-to-end task of IE has not been studied much. From the 120 papers included in this literature review, only 8 consider full IE. Experimental results show that performances are far from being usable in practice.
- Document-level IE. Most methods focus on sentence-level IE. Documents present unique challenges, such as long context, distant relations, and multi-hop reasoning, that are difficult to tackle in practice.
- Open-world IE. Most models restrict themselves to closed-world IE, where entity and relation types are known beforehand. Type discovery and auto-structuration also come with challenges and impacts on performances.

In the following chapters, we have tried to bring new contributions to these under-explored areas of IE.

III Linked-DocRED

Enhancing DocRED with Entity Linking to Evaluate End-To-End Document-Level Information Extraction

Training and evaluating information extraction models requires a dataset annotated with entities, coreferences, relations, and entity linking. However, existing datasets either lack entity linking labels, are too small, are not diverse enough, or are automatically annotated (without a strong guarantee of the correction of annotations).

In this chapter, we propose Linked-DocRED, the first manually annotated, large-scale, document-level IE dataset, to the best of our knowledge. We enhance the existing and widely-used DocRED dataset with entity linking labels that are generated thanks to a semi-automatic process that guarantees high-quality annotations. In particular, we use hyperlinks in Wikipedia articles to provide human-quality disambiguation candidates. We also propose a complete framework of metrics to benchmark end-to-end IE models by extending the work of Zaporojets et al. [21]. In particular, we define an entity-centric metric to evaluate entity linking and generalize the supervised metrics to work under open-world and unsupervised settings. Evaluating a supervised baseline shows promising results while highlighting the numerous challenges towards end-to-end information extraction.

Linked-DocRED, the source code for the entity linking, the baseline, and the metrics are distributed under an open-source license and can be downloaded from a public repository[1].

Most of the work described in this chapter, including text, figures, and tables, was presented at SIGIR'23 [32].

Contents

[1] Available at https://github.com/alteca/Linked-DocRED.

III.1 Introduction

Information Extraction (IE) aims to extract meaningful information from documents, that is, entities and relations between these entities, to build or complement a Knowledge Graph (KG). IE is a four-step process with 1. Named Entity Recognition (NER), 2. Coreference Resolution (CR), 3. Entity Linking (EL), and 4. Relation Extraction (RE).

Several datasets have been proposed to train and evaluate IE models. The most recent ones [1, 21, 203] focus on document-level information extraction, a more realistic, albeit more challenging scenario than sentence-level IE. Indeed, it raises new problems, such as the combinatorial explosion of relation extraction[2] or the difficulty of considering long documents' context. At the same time, documents offer opportunities; in particular, the longer and richer context can help extract more refined and precise information with less ambiguity.

However, quantitative and qualitative analyses show that none of the existing datasets is satisfactory for the end-to-end evaluation of IE models, covering the four NER, CR, RE, and EL steps. On the one hand, most datasets focus on the first three tasks, ignoring the last entity linking step [1, 204]. And yet entity linking is one of the most important steps, if not the most important, as it transforms ambiguous extracted triples (containing natural language) into structured and disambiguated nodes and relations. Without it, it is impracticable to build a knowledge graph, as it will contain many duplicated nodes or wrongfully merged ones. The question of ambiguity in natural language is indeed predominant: a surface form can refer to multiple entities (e.g., Georgia the Eastern Europe country, or Georgia in the U.S.), and an entity can be expressed with multiple surface forms (e.g., Anakin Skywalker and Darth Vader). Ignoring entity linking hides an important part of the complexity of extracting information.

[2]The number of relation candidates increases quadratically depending on the number of entities (as $\binom{n}{2} = \frac{n(n-1)}{2} \in \mathcal{O}(n^2)$).

On the other hand, datasets that provide entity linking annotations are either too small, not diverse enough, or automatically annotated [21, 203, 205, 206]. The shortfalls of these datasets are linked to the fact that high-quality entity linking annotations are generally expensive and time-consuming to produce due to the vastness of the search space. For instance, Wikidata[3], a reference knowledge base, contains more than 100 M entities.

Therefore, we propose Linked-DocRED, to the best of our knowledge, the first large-scale, manually labeled, document-level IE dataset that provides high-quality annotations for entities, coreferences, relations, and entity linking. Linked-DocRED aims to correct existing datasets' shortcomings and define a reproducible and more complete benchmark for the training and evaluation of end-to-end IE models. Instead of creating a dataset from scratch, we start from the widely-used DocRED dataset of Yao et al. [1]. DocRED is a general-domain dataset composed of 5,053 Wikipedia abstracts annotated with entities, coreferences, and relations. We enhance DocRED by labeling each entity with entity linking. Since DocRED documents are taken from Wikipedia articles, we propose to use wikilinks (internal Wikipedia hyperlinks) to generate entity linking annotations. It allows us to create a semi-automatic entity linking process that guarantees a human-quality annotation while being much faster and less expensive to implement. A thorough evaluation of the entity linking process shows the quality of our labeling. Our method can be replicated to other datasets based on Wikipedia articles, regardless of their language (e.g., HacRED [204]).

We also continue the work of Zaporojets et al. [21] by establishing a clear and coherent set of entity-centric metrics to evaluate the performance of an IE model. In particular, we define an entity-centric metric to assess entity linking and generalize existing metrics to work in an open-world and unsupervised setting. The evaluation of a baseline method based on recent approaches shows encouraging results. However, it demonstrates that this task is still a difficult challenge, particularly because of cascading errors during the successive steps of an IE pipeline. We hope that Linked-DocRED can facilitate the discovery of more performant IE models. To summarize our main contributions:

- We propose Linked-DocRED, the first large-scale, manually-labeled, document-level IE dataset built semi-automatically on top of the DocRED dataset. Linked-DocRED contains four times more entities and two times more relations than its closest competitor, DWIE [21] (section III.4).
- We propose a new entity linking method based on the alignment between DocRED documents and Wikipedia articles, providing high-quality labeling. This method can be applied to disambiguate other Wikipedia-based datasets (section III.3).
- We define a novel entity-centric metric to assess entity linking and generalize Zaporojets et al. [21] metrics to the open-world and unsupervised settings. Doing so provides a complete set of metrics to evaluate an IE model (section III.5).
- We adapt state-of-the-art approaches to provide a simple and reproducible baseline covering the four steps of IE, namely NER, CR, RE, and EL. The experimental results

[3] Available at https://www.wikidata.org.

are promising, with a large margin of progress, particularly for entity linking, which is subject to cascading errors at the end of the pipeline (section III.6).

III.2 Related Work

As reviewed in chapter II, recent papers often consider the NER, CR, and RE tasks of information extraction [22, 176, 177, 180, 207, 208], setting entity linking aside. Only a handful of papers [20, 44, 45] are exploring the end-to-end IE process. However, entity linking is critical, as it constitutes the bridge between extracted triples, which are ambiguous, and structured knowledge that downstream applications can use.

To train and evaluate IE models, numerous datasets have been proposed, covering a large spectrum of settings and applications:

- Some focus on general domain information (e.g., T-REx [205], DocRED [1], or HacRED [204]), others on very specific domains (scientific literature for SciERC [207], biomedicine for FewRel 2.0 [206], or BC5CDR [209]).
- Some are manually annotated (e.g., DocRED [1], FewRel [210] or HacRED [204]), others automatically generated such as T-REx [205] or NYT-10 [157].
- Some focus on sentences (e.g., FewRel [206, 210], or NYT-10 [157]) others on documents (DocRED [1], KnowledgeNet [203], or DWIE [21]).

We recall some characteristics of the major information extraction datasets in table III.1.

FewRel [206, 210] FewRel is large-scale, diverse (80 relation types), and annotated for the four tasks but does not contain documents. This lack of documents also explains the low number of coreferences compared to other datasets. Besides, FewRel does not contain new knowledge (all entities are already present in the knowledge base), simplifying the entity linking, as there are no unknown entities. It is thus not usable in practice for our scenario.

T-REx [205] It is the largest of the seven datasets, with around 4.6 M documents. It is not usable in our scenario, though, as the dataset was automatically labeled using IE models, which means there is no strong guarantee of the quality of annotations. Similarly to FewRel, it does not contain new knowledge (all entities are linked to Wikidata items). Nevertheless, it provides a huge source of distant-supervision, which can be beneficial during training (even though it is a lower-quality annotation).

KnowledgeNet [203] and BC5CDR [209] They contain documents and are annotated for the four tasks. Similarly to FewRel and T-REx, BC5CDR has no new knowledge (all entities are already in the knowledge base). This default is absent of KnowledgeNet, where 18 % of its entities do not exist in the KG. However, BC5CDR and KnowledgeNet are too small (see table III.1) and not diverse enough (with only 15 relations types for KnowledgeNet and 1 for BC5CDR), which raises questions regarding their representativeness for realistic IE scenarios.

Table III.1: Quantitative comparison between Linked-DocRED and widely-used IE datasets. Please refer to the legend below for additional details on the columns.

Dataset	Size		Entities		# Coreferences
	# Docs	# Tokens	# Entities	# Types	
FewRel [206, 210]	-	1,397 k	112 k	-	2 k
T-REx [205]	4,650 k	446,053 k	69,962 k	-	17,617 k
KnowledgeNet [203]	4 k	734 k	11 k	-	7 k
BC5CDR [209]	1.5 k	343 k	10 k	2	19 k
DWIE [21]	0.8 k	501 k	23 k	311	20 k
HacRED [204]	9.2 k	1,141 k	99 k	9	19 k
DocRED [1]	5.1 k	1,001 k	99 k	6	34 k
Linked-DocRED	5.1 k	1,001 k	95 k	6	38 k

Dataset	Entity Linking				Relations	
	# Linked		# New		# Instances	# Types
FewRel [206, 210]	112 k	(100 %)	0	(0 %)	56 k	80
T-REx [205]	69,962 k	(100 %)	0	(0 %)	208,774 k	642
KnowledgeNet [203]	9 k	(82 %)	1.9 k	(18 %)	13 k	15
BC5CDR [209]	10 k	(100 %)	0	(0 %)	48 k	1
DWIE [21]	13 k	(57 %)	10 k	(43 %)	22 k	65
HacRED [204]	-		-		68 k	26
DocRED [1]	-		-		50 k	96
Linked-DocRED	63 k	(67 %)	6.4 k	(7 %)	50 k	96

To understand the tables' columns:

- Size > # Docs. Number of documents.
- Size > # Tokens. Number of words and punctuations.
- Entities > # Entities. Number of entities ignoring coreferences.
- Entities > # Types. Number of entity types.
- # Coreferences. Number of coreference clusters.
- Entity Linking > # Linked. Number (secondary column percentage) of entities linked to a knowledge graph.
- Entity Linking > # New. Number (secondary column percentage) of entities that do not exist in the knowledge graph.
- Relations > # Instances. Number of relations.
- Relations > # Types. Number of relation types.

DWIE [21] Like KnowledgeNet and BC5CDR, DWIE contains documents labeled for the four tasks. It is more diverse and bigger, though, but still much smaller in terms of documents, entities, and relations than HacRED and DocRED. The analysis of the dataset's files suggests that entity linking was automatically labeled as each entity is associated with multiple candidates with eighteen-digit precision probabilities. As a result, it is not satisfactory for our purpose.

DocRED [1] and HacRED [204] They contain around two to five times more documents and annotations than the other manually annotated datasets, which makes them more suitable for training and evaluating IE models. Unfortunately, they are not annotated with entity linking.

Although several datasets have been proposed to evaluate IE models, none is satisfactory. Indeed, FewRel [206, 210] lacks documents and novel entities; T-REx [205] lacks manual annotations and novel entities; KnowledgeNet [203] and BC5CDR [209] are too small and not diverse enough; DWIE has automatic entity linking annotations [21]; and HacRED [204], and DocRED [1] lack annotations for entity linking. As a result, it motivates us to create a new dataset that would provide a complete and objective baseline to test and develop end-to-end IE models.

III.3 Dataset Generation

In this section, we describe the process we used to create Linked-DocRED. First, creating an IE dataset from scratch is an expensive endeavor as it requires annotating documents for entities, coreferences, relations, and entity linking. In particular, entity linking is very time-consuming due to the vastness of the search space (millions of candidates) and the ambiguity of natural language: an entity can have different surface forms, and the same surface form can refer to multiple entities (cf. Georgia presented in the introduction). At the same time, we notice that one existing dataset, DocRED [1], is almost adequate to train and evaluate an IE model, except for the lack of entity linking annotations. DocRED is also widely used and acknowledged for its quality as a benchmark, especially for document-level IE. Therefore, instead of creating a new dataset from the ground up, we propose to enhance DocRED with entity linking.

To create entity linking annotations, we do not want to rely on any automatic entity linker (for instance, DBpedia Spotlight [211]), as some competitor datasets (DWIE or T-REx) have implemented. Indeed, entity linkers would introduce biases. They are imperfect (in fact, even human annotation is imperfect), have advantages and drawbacks, and may wrongly link some entities. So, if an IE model uses the same entity linker for its predictions, it will reproduce the same behavior and obtain overstated results. The only valid choice for us is to rely on manual annotations to limit the introduction of bias in Linked-DocRED and ensure a fair comparison between different entity linking methods.

Our entity linking aims to link every entity of DocRED to a resource in Wikipedia (and we also provide the Wikidata ID associated with the Wikipedia resource). For the entities that do not exist in Wikipedia, we will assign them a unique identifier of the form `#DocRED-<id>#` (e.g., Ben Skywalker[4] in figure III.8). Providing two entity linking identifiers is beneficial:

[4]This entity is not in Wikipedia at the time we write this thesis.

> **Luke Skywalker** is a fictional character and the main protagonist of the original film trilogy of the *Star Wars* franchise created by George Lucas. The character, portrayed by Mark Hamill, is an important figure in the Rebel Alliance's struggle against the Galactic Empire. He is the twin brother of Rebellion leader Princess Leia Organa of Alderaan, a friend and brother-in-law of smuggler Han Solo, an apprentice to Jedi Masters Obi-Wan "Ben" Kenobi and Yoda, the son of fallen Jedi Anakin Skywalker (Darth Vader) and Queen of Naboo/Republic Senator Padmé Amidala and maternal uncle of Kylo Ren/Ben Solo. The now non-canon and discarded *Star Wars Legends* depicts him as a powerful Jedi Master, husband of Mara Jade, the father of Ben Skywalker and maternal uncle of Jaina, Jacen and Anakin Solo.
>
> In 2015, the character was selected by *Empire* magazine as the 50th greatest movie character of all time.[2] On their list of the *100 Greatest Fictional Characters*, Fandomania.com ranked the character at number 14.[3]

Figure III.1: Wikipedia abstract of Luke Skywalker with its wikilinks (in blue), as it was available on July 28, 2018[5]. It corresponds to instance 2,774 of DocRED/Linked-DocRED in figure III.8.

Wikipedia gives access to verbose and descriptive texts about the entity, and Wikidata to the interconnected structure of a knowledge graph.

A document in DocRED is a Wikipedia abstract, the first paragraphs of a Wikipedia article. If we take the instance presented in figure III.8, the document corresponds to the Wikipedia abstract of Luke Skywalker (displayed in figure III.1). The wikilinks in the Wikipedia abstract are interesting: they surround a term for which they indicate the URL of the Wikipedia article defining it. It is a form of entity linking between entities in the abstract and Wikipedia resources, as there is only one Wikipedia article (one URL) defining a specific concept. To be more precise, as Wikipedia contributors manually edit these wikilinks, this can be considered as a form of manual entity linking. In brief, it is an ideal source of high-quality entity linking for DocRED.

Besides, we note that there is a direct mapping (same sentence, same position) between a lot of DocRED entities in the document and wikilinks in the corresponding Wikipedia article (e.g., "Star Wars", "George Lucas", "Mark Hamill", "Padmé Amidala", or "Galactic Empire" in figures III.1 and III.8). Using these wikilinks with this very strict mapping is the basic idea we developed for our semi-automatic, high-quality entity linking.

The general process we used to annotate DocRED with entity linking is presented in figure III.2. The main step is mapping entities with wikilinks, the second module of figure III.2 (Wikilinks Alignment, presented in section III.3.2). It is insufficient to fully disambiguate our dataset, which explains the three steps following it. In the next sections, we will describe each

[5]Taken from https://en.wikipedia.org/w/index.php?title=Luke_Skywalker&oldid=876119706 (visited on 2024-07-08).

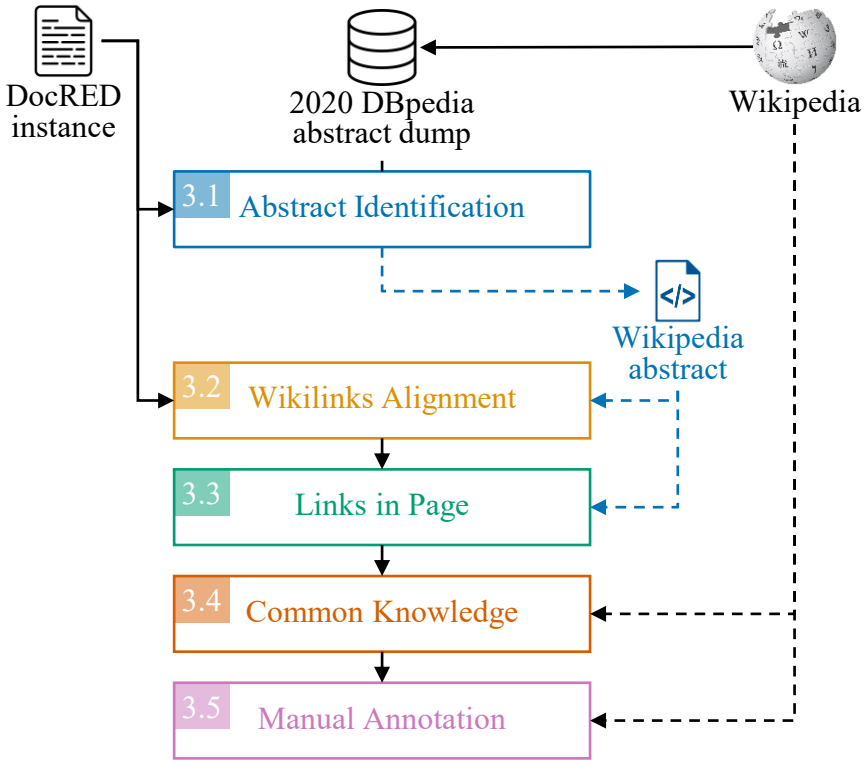


Figure III.2: Architecture of the semi-automatic entity linking process implemented to disambiguate Linked-DocRED.

constituent in the disambiguation process for a DocRED document.

NUM and TIME Entities

Within DocRED, 25,171 entities (26.6 %) are numerals (NUM) or temporal (TIME) entities. In a knowledge graph such as Wikidata or DBpedia[6], these entities are not considered resources (associated with a unique URI) but literals, which are not disambiguated. Although the disambiguation of dates and numbers could be interesting, we apply the same rule for Linked-DocRED and create a particular identifier `#ignored#` to indicate no disambiguation for NUM and TIME entities.

III.3.1 Wikipedia Abstract Identification

To access the wikilinks and map them to our entities, we first need to get the Wikipedia article associated with our DocRED document (the first step in figure III.2). Although DocRED does not contain the URL of the source Wikipedia page, we can access the article title and the abstract. The solution we implemented is a full-text search on the title and abstract to find the most similar Wikipedia article.

Another aspect to consider is that DocRED was published in 2019, meaning that many Wikipedia pages have been modified since, which can lead to poor results with full-text searches.

[6]Available at https://www.dbpedia.org.

To mitigate this issue, we downloaded the 2020-01 DBpedia abstracts dump[7], which is the oldest available this day. We have also tested with Wikipedia dumps but found them of lower quality than DBpedia: some abstracts were truncated, and others contained abnormal characters. From the DBpedia dump, 5.6 M Wikipedia abstracts were indexed in ElasticSearch[8]. As a side note, the dump does not contain all Wikipedia articles: when we write the thesis, there are around 6.8 M Wikipedia articles.

For a given DocRED document, we then perform a full-text search comparing the instance text to the abstracts in ElasticSearch to identify the Wikipedia abstract most similar to our document. Internally, ElasticSearch uses inverted full-text indices provided by Apache Lucene[9] and the BM25 metric [212] to perform its search (weighted bag-of-words). This setup is efficient and fast in returning good Wikipedia candidates, but it does not consider word order in the Wikipedia article. In practice, the word order is critical to determine the correct Wikipedia abstract. To have the best confidence possible, we propose to rank the candidates using a similarity metric based on the Levenshtein distance [213]. The Levenshtein distance measures the number of character modifications (insertion, deletion, modification) to transform the first string into the second. It is defined recursively[10] as follows:

$$\mathrm{lev}(\boldsymbol{d}_1, \boldsymbol{d}_2) = \begin{cases} |\boldsymbol{d}_1| & \text{if } |\boldsymbol{d}_2| = 0, \\ |\boldsymbol{d}_2| & \text{if } |\boldsymbol{d}_1| = 0, \\ \mathrm{lev}(\mathrm{tail}(|\boldsymbol{d}_1|), \mathrm{tail}(|\boldsymbol{d}_2|)) & \text{if } \mathrm{head}(|\boldsymbol{d}_1|) = \mathrm{head}(|\boldsymbol{d}_2|), \\ 1 + \min \begin{cases} \mathrm{lev}(\mathrm{tail}(|\boldsymbol{d}_1|), \boldsymbol{d}_2) \\ \mathrm{lev}(|\boldsymbol{d}_1|, \mathrm{tail}(\boldsymbol{d}_2)) \\ \mathrm{lev}(\mathrm{tail}(|\boldsymbol{d}_1|), \mathrm{tail}(|\boldsymbol{d}_2|)) \end{cases} & \text{otherwise,} \end{cases} \tag{III.1}$$

with $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ the two documents to be compared, $|\boldsymbol{d}_1|$ the length (number of characters) of $\boldsymbol{d}_1$, $\mathrm{head}(\boldsymbol{d}_1) = \boldsymbol{d}_1[0]$ the first character of $\boldsymbol{d}_1$, and $\mathrm{head}(\boldsymbol{d}_1) = \boldsymbol{d}_1[1{:}]$ the second and following characters of $\boldsymbol{d}_1$. We derive a Levenshtein-based similarity metric defined between $[0, 1]$ as:

$$\mathrm{lev}_{\mathrm{sim}}(\boldsymbol{d}_1, \boldsymbol{d}_2) = 1 - \frac{\mathrm{lev}(\boldsymbol{d}_1, \boldsymbol{d}_2)}{\max(|\boldsymbol{d}_1|, |\boldsymbol{d}_2|)}. \tag{III.2}$$

This similarity metric ranks precisely the candidates: logically, the DocRED document and the correct Wikipedia candidate are the closest regarding Levenshtein distance. However, it cannot determine whether the first Wikipedia candidate is the right article. Indeed, as we have said previously, our DBpedia dump is incomplete: it does not contain every Wikipedia abstract, which means that some DocRED documents cannot be found. We propose determining a threshold with $\mathrm{lev}_{\mathrm{sim}}$ to filter those instances.

[7] Available at https://databus.dbpedia.org/dbpedia/text/long-abstracts.

[8] Available at https://www.elastic.co/elasticsearch.

[9] Available at https://lucene.apache.org.

[10] We provide the recursive formulation due to its clarity. In practice, dynamic programming iterative algorithms are preferred as they have a lower quadratic complexity. In particular, the Needleman-Wunsch algorithm [132] defined in figure III.4 can be used to compute the Levenshtein distance.

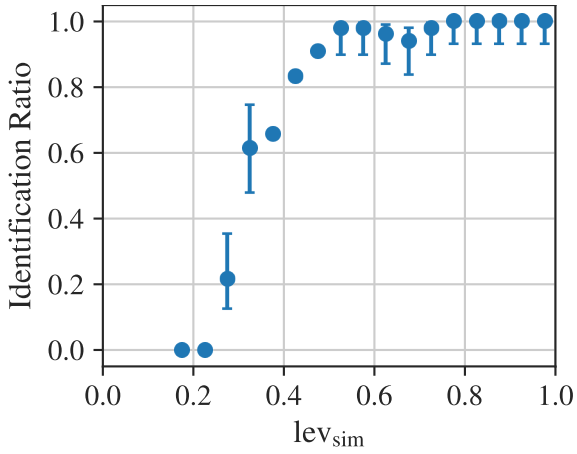


Figure III.3: Evolution of the correct Wikipedia identification ratio depending on lev_{sim} between the DocRED instance and the candidate Wikipedia abstract. The Wikipedia identification ratio is computed based on the manual annotation of 1,000 pairs of DocRED instance and Wikipedia abstract stratified in 20 bins of size 0.05 lev_{sim}. We show the Wilson confidence interval with $\alpha = 0.05$. No confidence interval is displayed if the bin contains less than 50 pairs.

To do that, we select a sample of 1,000 DocRED documents with their first Wikipedia candidate, stratified with lev_{sim} (20 bins of size 0.05, containing 50 instances each), and we manually determine whether the Wikipedia candidate is correct or not. The results are shown in figure III.3. For each bin, we also compute the confidence interval for the proportion with a standard significance level $\alpha = 0.05$, using the Wilson approximation [214]. Indeed, we have few instances per bin coupled with ratios close to 0 and 1; therefore, the central limit theorem approximation does not hold, and the usual Wald confidence interval becomes unreliable. On the contrary, the Wilson score interval can be used with small samples and extreme proportions. It is defined as follows:

$$\text{CIW}_\alpha = \left[\frac{1}{1 + \frac{z_\alpha^2}{n}} \left(\hat{p} + \frac{z_\alpha^2}{2n} \pm \frac{z_\alpha}{2n} \sqrt{4n\hat{p}(1 - \hat{p} + z_\alpha^2)} \right) \right], \tag{III.3}$$

with n the size of the sample, $\hat{p}$ the observed proportion, and z_α the $1 - \frac{\alpha}{2}$ quantile of a standard normal distribution[11]. In practice, some bins contain less than 50 elements (e.g., $]0.15, 0.20]$, $]0.35, 0.40]$, or $]0.40, 0.45]$), in which case we annotate all instances, so there is no confidence interval.

We notice that for $\text{lev}_{\text{sim}} > 0.5$, the proportion of correctly identified Wikipedia articles is close to 1 (above 0.95). Therefore, we propose to set our threshold at $\text{lev}_{\text{sim}} > 0.5$ and manually check DocRED documents with $\text{lev}_{\text{sim}} \leq 0.5$. Using this threshold, we automatically identify the Wikipedia article for 4,694 documents (93 %). We manually determine the Wikipedia abstract for the remaining 357 documents. We could not find the Wikipedia article for 23 instances; we think these articles have been completely removed from Wikipedia. These DocRED instances will be manually disambiguated.

[11] With $\alpha = 0.05$, we have $z_\alpha \approx 1.96$.

Finally, once the correct Wikipedia abstract is identified, we download the HTML code of the corresponding Wikipedia article. We select the page revision ID to be before 2018-12-31 to minimize the differences between the Wikipedia page and the DocRED instance. As a side note, the target date was arbitrarily fixed at 2018-12-31, as we could not precisely determine the timestamp of the extraction made by Yao et al. In hindsight, we found that their extraction was anterior to December 2018 and probably close to summer 2018. In practice, this six-month offset did not impact the results[12].

III.3.2 Wikilinks Alignment

In this module (second step in figure III.2), we implement the mapping between the DocRED document's entities and wikilinks in the Wikipedia article we downloaded in the previous section. We want to find direct intersections (same sentence and position) between entities in our DocRED instance and wikilinks in the Wikipedia article to maximize the confidence in the annotations. To do that, we need to align our DocRED text precisely with the Wikipedia abstract. The problem is that there are minor differences between the two texts, which make this step nontrivial. These small discrepancies come from the preprocessing applied by Yao et al. on DocRED instances that remove Cyrillic, Arabic, and Asiatic characters and some parts of the abstract, and by the small revision modifications on Wikipedia articles (as we have not been able to perfectly identify the date when Yao et al. retrieved the articles).

To overcome this difficulty, we propose to use the Needleman-Wunsch algorithm [132], which was initially proposed to optimally align two nearly-identical DNA sequences, minimizing insertions, deletions, and substitutions of nucleotides. This algorithm is easily generalizable to string alignment by replacing the notion of nucleotides with characters. It is equivalent to the Levenshtein distance, with the complement that it not only returns the editing distance but also the list of modifications to transform one string in the other. It produces a translation table to convert a character position in the DocRED instance to a position in the Wikipedia article.

The Needleman-Wunsch algorithm relies on dynamic programming principles: it solves small alignment problems that incrementally lead to the optimal global solution. The algorithm is presented in figure III.4. A 2-D array ($\boldsymbol{F}$) is used to store the optimal alignment solutions of substrings of the DocRED instance $\boldsymbol{d}_D$ (first dimension) and the Wikipedia article $\boldsymbol{d}_W$ (second dimension). The substrings start at position 0 and end at position i for DocRED and j for Wikipedia. Another 2-D array ($\boldsymbol{A}$) with a similar shape as $\boldsymbol{F}$ is used to store the operations (insertion, deletion, substitution) selected by the optimal alignment. Insertions, deletions, and substitutions are associated with penalties: $s_{INS} = 1$ for insertions and deletions[13], and $s_{SUB} = 1$ for substitutions. The algorithm is composed of three steps.

Initialization $\boldsymbol{F}$ represents the editing distance of the candidate alignment (error). Solutions with low distances are better than those with high distances. Insertions, deletions, and

[12] It likely increased the amount of manual annotation required.

[13] Insertion and deletions are the same operations but with a different point of view (deletion from the point of view of $\boldsymbol{d}_D$ is equivalent to insertion from the point of view of $\boldsymbol{d}_W$).

function *Needleman-Wunsch($\boldsymbol{d}_D$, $\boldsymbol{d}_W$)*
begin
  Data: $\boldsymbol{d}_D$ text of the DocRED document, $\boldsymbol{d}_W$ text of the Wikipedia article
  Result: $\boldsymbol{M}$ the list of modifications to transform $\boldsymbol{d}_D$ into $\boldsymbol{d}_W$
  $s_{INS}, s_{SUB}, s_{MATCH} \leftarrow 1, 1, 0$
  $\boldsymbol{F}, \boldsymbol{A} \leftarrow \text{Array}(|\boldsymbol{d}_D| + 1, |\boldsymbol{d}_W| + 1), \text{Array}(|\boldsymbol{d}_D| + 1, |\boldsymbol{d}_W| + 1)$
  `/* Initialization */`
  $\boldsymbol{F}[0, 0] \leftarrow 0$
  for $i \leftarrow 1$ **to** $|\boldsymbol{d}_D| + 1$ **do**
    $\boldsymbol{F}[i, 0] \leftarrow \boldsymbol{F}[i - 1, 0] + s_{INS}$
    $\boldsymbol{A}[i, 0] \leftarrow$ "DEL"
  end
  for $j \leftarrow 1$ **to** $|\boldsymbol{d}_W| + 1$ **do**
    $\boldsymbol{F}[0, j] \leftarrow \boldsymbol{F}[0, j - 1] + s_{INS}$
    $\boldsymbol{A}[0, j] \leftarrow$ "INS"
  end
  `/* Optimal Alignment */`
  for $i \leftarrow 1$ **to** $|\boldsymbol{d}_D| + 1$, $j \leftarrow 1$ **to** $|\boldsymbol{d}_W| + 1$ **do**
    $ins \leftarrow \boldsymbol{F}[i, j - 1] + s_{INS}$
    $del \leftarrow \boldsymbol{F}[i - 1, j] + s_{INS}$
    $match \leftarrow \boldsymbol{F}[i - 1, j - 1] +$ **if** $\boldsymbol{d}_D[i] = \boldsymbol{d}_W[j]$ **then** s_{MATCH} **else** s_{SUB}
    $\boldsymbol{F}[i, j] \leftarrow \min(ins, del, match)$
    $\boldsymbol{A}[i, j] \leftarrow$ ("INS", "DEL", "MATCH/SUB")[argmin($ins, del, match$)]
  end
  `/* Backtracking */`
  $\boldsymbol{M} \leftarrow []$
  while $i > 0$ **do**
    if $\boldsymbol{A}[i, j] =$ *"DEL"* **then**
      prepend "DEL" to $\boldsymbol{M}$
      $j \leftarrow j - 1$
    else if $\boldsymbol{A}[i, j] =$ *"INS"* **then**
      prepend "INS" to $\boldsymbol{M}$
      $i \leftarrow i - 1$
    else
      prepend **if** $\boldsymbol{d}_D[i] = \boldsymbol{d}_W[j]$ **then** "MATCH" **else** "SUB" to $\boldsymbol{M}$
      $i, j \leftarrow i - 1, j - 1$
  end
end

Figure III.4: Pseudocode of the Needleman-Wunsch algorithm employed to align the text of the DocRED instance with the Wikipedia article.

substitutions come with a penalty compared to a match. Thus, the idea is to minimize the penalties. We initialize $\boldsymbol{F}[0,0] = 0$ (aligning "" with "" requires no operation), $\boldsymbol{F}[i,0] = i \cdot s_{INS}$ (the only possibility to translate $\boldsymbol{d}_D$ into "" is to do deletions), and $\boldsymbol{F}[0,j] = j \cdot s_{INS}$ (the only possibility to translate "" into $\boldsymbol{d}_W$ is to do insertions).

Optimal Alignment $\boldsymbol{F}$ is then filled iteratively. At each step, three editing distances are computed:

- Insertion. The distance if we do an insertion operation at this step.
- Deletion. The distance if we do a deletion operation at this step.
- Match/Substitution. The distance if we do a match or a substitution. If the two characters are equal, we have a match; otherwise, we have a substitution.

These three distances can be easily calculated using the previously filled values[14] in $\boldsymbol{F}$ (insertion uses $\boldsymbol{F}[i, j-1]$, deletion $\boldsymbol{F}[i-1, j]$, and match/substitution $\boldsymbol{F}[i-1, j-1]$). Then, the operation that minimizes the editing distance is chosen and stored in $\boldsymbol{A}$.

Backtracking Once $\boldsymbol{F}$ is filled, the list of modifications between $\boldsymbol{d}_D$ and $\boldsymbol{d}_W$ are computed by backtracking $\boldsymbol{A}$.

With the translation table built using the Needleman-Wunsch algorithm, it is simple to compute intersections between surface forms of entities as annotated in DocRED and wikilinks and thus generate entity linking candidates.

However, we have no guarantee of the quality of the proposed candidates. Intuitively, if the intersection is exact, the entity linking should be accurate. It becomes more difficult with a partial intersection. For instance, "Columbia University in the City of New York" is the same as "Columbia University", but "Columbia" is not the same entity as "Columbia University". A simplistic measure is to keep only exact intersections, but we would discard many good disambiguations.

Instead, we propose to evaluate the impact of the quality of the intersection on the disambiguation. To do this, we apply a method similar to that of section III.3.1. We first compute lev_{sim}, defined equation (III.2), between the DocRED entity text and the matched wikilink, which allows us to quantify the intersection quality. We then select a sample of 1,000 entities and their matched wikilinks, stratified on lev_{sim} (with 20 bins of 0.05), and manually determine whether the entity linking is correct. We also compute a Wilson confidence interval for each bin for a proportion with $\alpha = 0.05$. The results are shown in figure III.5.

We can see three regimes:

- $\text{lev}_{\text{sim}} < 0.35$. Few entities are correctly disambiguated, which is logical given that the entity and the wikilink are dissimilar.

[14] As a side note, it is possible because the algorithm only allows modifications that do not impact "future" characters (such as permutations).

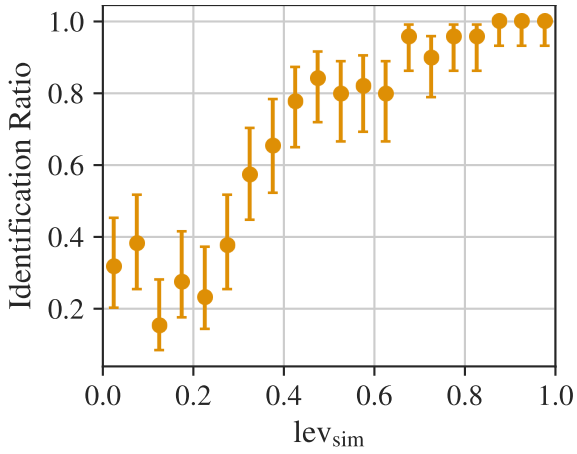


Figure III.5: Evolution of the correct entity linking identification ratio depending on lev_{sim} between the entity and the candidate wikilink. The correct entity linking identification ratio was computed by manually annotating 1,000 pairs of entity and candidate wikilink stratified in 20 bins of size 0.05 lev_{sim}. We show the Wilson confidence interval with $\alpha = 0.05$.

- $\text{lev}_{\text{sim}} \in [0.35, 0.75]$. Entities and wikilinks are relatively similar, but the probability of wrong entity linking is still high.

- $\text{lev}_{\text{sim}} > 0.75$. The proportion is close to 1 (0.984): of the 250 annotated pairs, only four are wrongly linked.

Considering this figure III.5, we keep only entity linking candidates with the highest correct entity linking identification ratio: $\text{lev}_{\text{sim}} > 0.75$. By doing so, we disambiguate 40,826 entities of DocRED (43.3 %) as shown in figure III.7.

This module provides entity linking annotations with high confidence, as we are strict with intersections and textual similarity and rely on manual annotations of Wikipedia contributors.

III.3.3 Links in Page

In a Wikipedia article, the first mention of an entity is associated with a wikilink, while the following, most often, are not. As a result, the entity may be disambiguated in the Wikipedia article but not in the specific span of text we are considering. A workaround is to check if a wikilink on the Wikipedia page has the same surface form as the entity we are disambiguating (the third step in figure III.2). We also have to check for ambiguity: if there are multiple wikilinks with the same surface form but different target articles, we cannot choose the correct one with this module. Using this module, we disambiguate 6,741 additional entities (7.1 %).

This approach is of lower quality compared to wikilinks alignment. However, we are strict on selecting wikilinks: we allow only an exact match between the wikilink and the surface form of the entity.

III.3.4 Common Knowledge

When analyzing the remaining undisambiguated entities, we notice that some of them are very common: famous persons (e.g., Bill Gates, Barack Obama), well-known companies (Facebook, Apple, ...), or common-knowledge places (United States, Spain, Paris, New York, etc.). These entities are so famous that they are not associated with wikilinks, as it is supposed that everyone knows them already.

To add this notion of common knowledge (the fourth step in figure III.2), we select the entities mentioned at least three times in the dataset and manually annotate them. We take particular care in detecting entities with ambiguities. For instance, *French* can refer to France, the French language, or the French people, and *Georgia* points to the eastern European country or the U.S. state. The ambiguity about French can be solved by looking at the types: France is a location (LOC), the French language is classified as miscellaneous (MISC), and French people is identified as an organization (ORG). However, the only possibility for Georgia is to label each instance manually (see next section). After this filtering step, we annotate around 1,000 entities, which leads to the disambiguation of 7,684 more entities (8.1 %).

We estimate the quality of the entity linking to be as good as links in page, as the two processes are similar.

III.3.5 Manual Annotation

We manually annotate the remaining 14,125 entities to guarantee a high-quality entity linking. Among these entities, we expect to be able to disambiguate the majority, but we also anticipate encountering entities that are not present in Wikipedia. To facilitate the labeling process, we designed an interface with Label Studio[15] (displayed in figure III.6).

The annotation is done document by document. The annotators must label every remaining entity (three per document on average). To help them, a list of five candidates per entity is provided from which they can choose. These candidates are determined by searching on Google[16] using the surface form and filtering to keep only Wikipedia results[17]. They can also manually enter a Wikipedia URL or a coreference with another entity in the document. Finally, they can indicate that the entity does not have a Wikipedia page (new knowledge).

A single annotator labeled all the entities to ensure maximal coherence in the entity linking scheme. During the manual annotation, he identified 523 errors in the dataset[18]: 361 entities were wrongly typed, 148 mentions needed to be corrected (the boundaries were wrong), and 14 mentions were not entities. These errors have been corrected.

Inter-Annotator Agreement To better understand the quality of the manual annotation, we selected a sample of 1,018 entities, and three annotators disambiguated them to check if the

[15]Available at https://labelstud.io.

[16]Available at https://www.google.com.

[17]We have also tried using the internal search engine of Wikipedia or Wikidata, but they achieve lower recall than Google.

[18]This error identification step is not exhaustive.

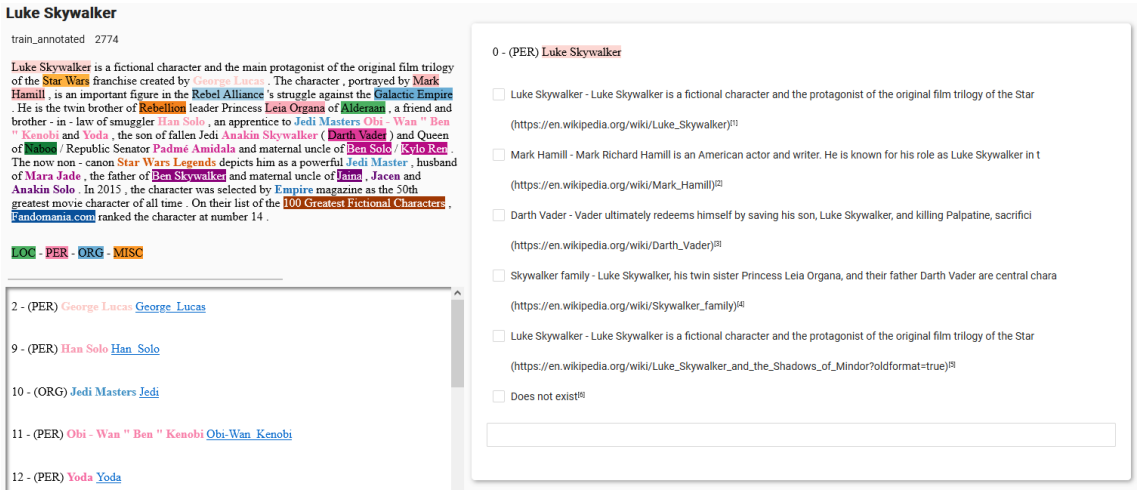


Figure III.6: GUI developed with Label Studio to manually disambiguate the remaining 14,125 entities in Linked-DocRED. On the upper left, the text of the instance is displayed (it corresponds to instance 2,774 shown in figure III.1 and figure III.8). The entities to disambiguate are highlighted, and the already linked ones have colored characters. Below is recalled each linked entity with its Wikipedia resource. On the right, the entity to disambiguate is displayed, with five candidates provided. The annotator can choose one, indicate a coreference in the text field, enter a manual wikilink, or signify that the entity does not exist in Wikipedia.

entity linkings were similar. On this sample, we compute the Cohen's kappa coefficient [215] to qualify the inter-annotator agreement and obtain

$$\kappa_{EL} = 0.679.$$

This κ_{EL} score shows a strong inter-annotator agreement, especially considering the diversity of Wikipedia resources (more than 6.8 M articles in Wikipedia). Looking more precisely at the disagreements, we notice that for 30 % of them, one annotator indicated that the entity does not exist in Wikipedia, while the other was able to find it. It shows the difficulty of being exhaustive in the search for a Wikipedia resource. If we correct these disagreements, we obtain a $\kappa_{EL} = 0.816$, indicating a strong agreement between annotators.

Overall, this inter-annotator agreement analysis exhibits high-quality annotations. The main weakness is the complexity of determining with certainty that an entity does not exist in Wikipedia. As a result, in the final dataset, we distinguish a manual annotation leading to a Wikipedia resource from a manual annotation leading to "does not exist." If the entity is considered new, we provide a unique entity linking identifier of the form `#DocRED-<id>#`. As the confidence is lower in this case, we provide the list of candidates that the annotators refused, as they are candidates that an entity linker can easily predict. We are sure that these candidates are wrong.

This five-step process allows us to label all entities in Linked-DocRED. The participation of all methods in the disambiguation can be seen in figure III.7. To find the Wikidata ID for each disambiguated entity, we use the metadata of the Wikipedia resource (the property `wikibase_item`).

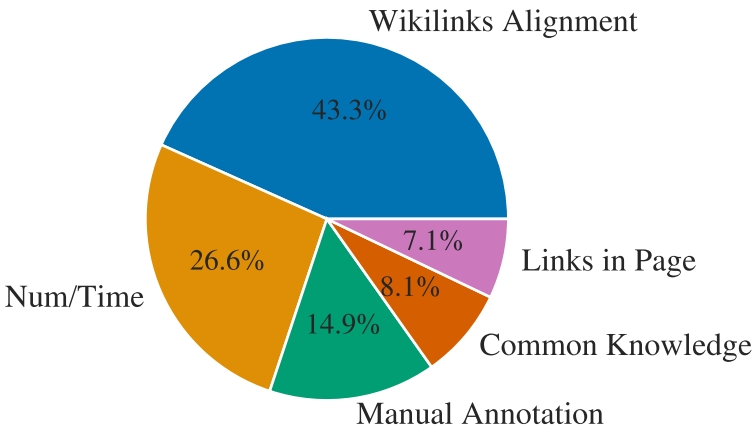


Figure III.7: Modules used to disambiguate the 94,547 entities of Linked-DocRED (see section III.3 for the details).

III.4 Dataset

As we have said earlier, Linked-DocRED comprises Wikipedia abstracts annotated with entities, coreferences, relations, and entity linking. The main statistics of the dataset are shown in the last line of table III.1.

The instance 2,774 of the train split is shown in figure III.8. The entities in the document are highlighted (pink for PER, orange for MISC, blue for ORG, green for LOC, grey for TIME, and brown for NUM). Two examples of entities are displayed below the document, with their mentions and the Wikipedia resource determined during entity linking. “Ben Skywalker” does not exist in Wikipedia; therefore, it is associated with the unique id `#DocRED-6032#`. Two examples of relations are also displayed. Finally, at the bottom, a small part of the knowledge graph representing the knowledge contained in the document is shown. In particular, we can see entities and relations that do not exist in Wikipedia or Wikidata (related to the node `#DocRED-6032#`).

III.4.1 Entities, Coreferences, Relations

We are using the entities, coreferences, and relations labels of DocRED; therefore, we recall the annotation process implemented by Yao et al. [1].

Entities & Coreferences Entities are automatically extracted and typed using spaCy[19]. To generate coreferences candidates, the entities are linked to Wikidata, with two basic approaches: 1. exact match between the surface form and a Wikidata entity label, or 2. using the TagMe entity linker [216]. As a side note, this primitive entity linking is not retained in their published dataset because its objective is not to be precise but to generate coreference and relations candidates. The entities and coreferences candidates are then corrected and complemented by human annotators.

[19]Available at https://spacy.io.

Luke Skywalker *(train, 2774)*

[0] Luke Skywalker is a fictional character and the main protagonist of the original film trilogy of the Star Wars franchise created by George Lucas. [1] The character, portrayed by Mark Hamill, is an important figure in the Rebel Alliance's struggle against the Galactic Empire. [2] He is the twin brother of Rebellion leader Princess Leia Organa of Alderaan, a friend and brother-in-law of smuggler Han Solo, an apprentice to Jedi Masters Obi-Wan "Ben" Kenobi and Yoda, the son of fallen Jedi Anakin Skywalker (Darth Vader) and Queen of Naboo / Republic Senator Padmé Amidala and maternal uncle of Ben Solo / Kylo Ren. [3] The now non-canon Star Wars Legends depicts him as a powerful Jedi Master, husband of Mara Jade, the father of Ben Skywalker and maternal uncle of Jaina, Jacen and Anakin Solo. [4] In 2015, the character was selected by Empire magazine as the 50th greatest movie character of all time. [5] On their list of the 100 Greatest Fictional Characters, Fandomania.com ranked the character at number 14.

ID 12	*Type* PER	*ID* 18	*Type* PER
Mentions	Anakin Skywalker, Darth Vader	*Mentions*	Ben Skywalker
Resource	`Darth_Vader (Q12206942)`	*Resource*	`#DocRED-6032#`

Head	0 (Luke Skywalker)	*Head*	18 (Ben Skywalker)
Tail	1 (Star Wars)	*Tail*	2 (George Lucas)
Relation	`present_in_work`	*Relation*	`creator`
Evidence	0	*Evidence*	0, 3

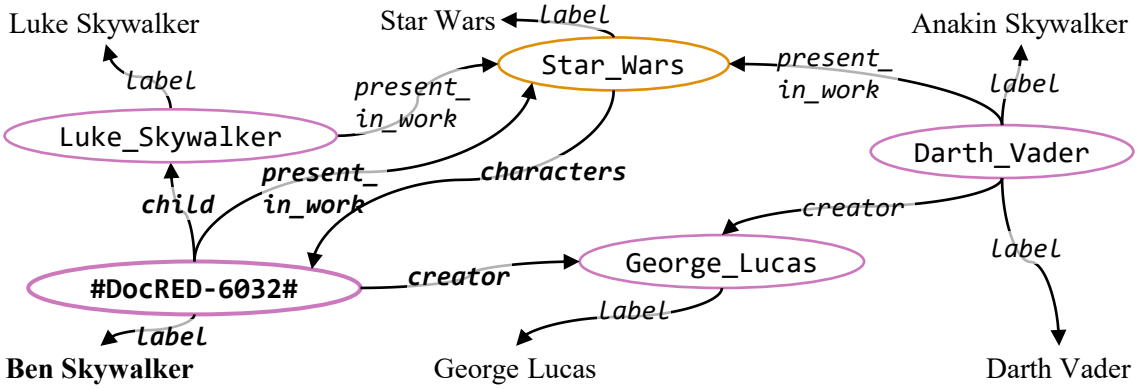


Figure III.8: Example instance of Linked-DocRED. From top to bottom: text of a document with highlighted entities; two examples of extracted entities with their Wikipedia and Wikidata resources, Ben Skywalker has no corresponding resource in Wikipedia; two examples of relations; and small part of the knowledge graph built from the entities and relations of the document.

Relations Using the basic entity linking, candidate relations are generated under the distant-supervision setting. Distant-supervision implies that if two entities, linked by a relation ***r*** in a knowledge graph, appear in the same document, then they express the relation ***r*** in the document. Other candidates are generated using RE models (not explained by Yao et al. [1]). The candidates are validated and supplemented by the annotators. Besides, annotators also indicate the sentences that support the existence of the relation in the document (evidence in figure III.8).

As we can see in table III.1, our entity linking does not impact the annotation of relations of DocRED, as the statistics of Linked-DocRED are identical to DocRED. However, it modifies coreferences (and entities indirectly): we identify 4,013 new coreferences that were not detected in DocRED. For example, in the instance 2,774 of the train split (see figure III.8), "Darth Vader" and "Anakin Skywalker" were not identified as coreferences.

III.4.2 Entity Linking

The entity linking annotation process is described in section III.3. To sum up, we rely on human annotations elicited by a semi-automatic process, as shown in figure III.2: 1–3. we map entities with wikilinks to benefit from Wikipedia contributor's annotations, 4. we use common knowledge (that was manually annotated), and 5. we manually label the remaining entities. This process leads to the disambiguation of every entity in Linked-DocRED. As we see in table III.1, 67 % of the entities are associated with a Wikipedia page and a Wikidata resource, and 7 % are identified as new resources unknown in Wikipedia. The remaining 26 % entities are numerals or temporal data not disambiguated, following Wikidata's and DBpedia's schemes.

For each entity in Linked-DocRED, we provide the following:

- `wikipedia_resource`: the identifier of the Wikipedia page, for instance `Darth_Vader` for entity 12 in figure III.8.
  If the entity is ignored (NUM or TIME), we have instead `#ignored#`.
  If the entity is new (unknown in Wikipedia), a unique identifier is provided of the form `#DocRED-<id>#`, for example, `#DocRED-6032#` for entity 18 in figure III.8.

- `wikidata_resource`: the identifier of the Wikidata entity, for instance `Q12206942` for entity 12 in figure III.8.

- `wikipedia_not_resource`: in the case of a new entity (unknown in Wikipedia), we provide the list of candidates that the annotator refused. They can be used to check that an entity linker is not predicting them.

- `method`: the method used to disambiguate this entity (see figure III.7).

- `confidence`: a confidence value from three choices: A, B, C.

Indeed, each entity linking in Linked-DocRED is associated with a confidence indicator. We define three possible classes:

Table III.2: Proportion of Linked-DocRED entities associated with each confidence indicator and estimation of the correct entity linking probability on a sample of 1,000 entities. The Wilson confidence interval with $\alpha = 0.05$ is printed in parenthesis. We do not estimate the correct entity linking probability for the C indicator, as it is difficult to ensure that an entity does not exist in Wikipedia due to the vastness of the search space.

Confidence Indicator	A	B	C
Proportion in Linked-DocRED	78.0 %	15.2 %	6.8 %
Correct entity linking probability	0.979(9)	0.950(14)	-

- A (very high confidence). The entity was linked using wikilinks alignment or manual annotation, or it was ignored (NUM and TIME).
- B (high confidence). The entity was linked using links in page or common knowledge.
- C (medium confidence). The annotator indicated that the entity does not exist in Wikipedia.

These indicators give a qualitative assessment of the quality of the disambiguation. To give a quantitative estimation of the probability of correct entity linking, we selected a sample of 1,000 entities for indicator A and 1,000 entities for indicator B. A human annotator manually checked these entities to determine whether the entity linking was correct. It allows us to estimate the probability of correct entity linking. We also compute a Wilson confidence interval for the proportion with $\alpha = 0.05$. We provide no estimation for indicator C as it is complicated to be sure that an entity does not exist in Wikipedia. The results are shown in table III.2.

The probabilities are close to 1 for indicators A and B, demonstrating the entity linking quality of Linked-DocRED. We note that the probability is a little higher for indicator A. Besides, we notice that 78 % of Linked-DocRED entities are scored as A, that is, with the highest confidence. Overall, the confidence is excellent throughout the whole dataset.

III.4.3 Linked-Re-DocRED

Concurrently to our work with Linked-DocRED, Tan et al. [217] found that DocRED was incomplete, notably regarding relation annotations. They addressed this shortcoming by complementing DocRED instances with the missing relation labels. They more than doubled the number of relations: they labeled 120 k relations compared to the initial 50 k of DocRED. They call this complemented dataset Re-DocRED. This results in increased performances for the baselines (with an F1 score gain of 13 %).

As this work complements Linked-DocRED, we incorporate their supplementary relation annotations in Linked-DocRED to form the Linked-Re-DocRED dataset (also available in our repository). In the evaluation part, we do not include the results on Linked-Re-DocRED, as we were unaware of this new dataset when they were run.

III.5 Entity-Centric Metrics to Evaluate Information Extraction

Metrics to evaluate an end-to-end IE model are a complex subject due to the existence of two points of view: mentions (low-level) and entities (higher-level). Most of the extraction is done with entities in mind, so evaluating the model from the entity perspective makes sense. However, comparing a ground truth entity with a predicted one is nontrivial because they can contain different mentions (no perfect intersection between true and predicted mentions) or nearly identical ones but not equal (differences in boundaries, for example).

III.5.1 Named Entity Recognition (Mention F1)

NER is the only module working with entity mentions. Similarly to previous works (Zhong et al. [113] among others), we consider a predicted mention to be correct if its boundaries are the same as the ones of a ground truth mention. Thus, we define true positives (TP), false positives (FP), and false negatives (FN) as follows:

$$\hat{\boldsymbol{m}} = \boldsymbol{m} \iff \text{start}(\hat{\boldsymbol{m}}) = \text{start}(\boldsymbol{m}) \wedge \text{end}(\hat{\boldsymbol{m}}) = \text{end}(\boldsymbol{m}), \tag{III.4}$$

$$\text{TP}_{NER} = \sum_{\hat{\boldsymbol{m}}} \sum_{\boldsymbol{m}} \mathbb{1}_{\hat{\boldsymbol{m}}=\boldsymbol{m}}, \tag{III.5}$$

$$\text{FP}_{NER} = \sum_{\hat{\boldsymbol{m}}} \mathbb{1}_{\neg\exists \boldsymbol{m} \text{ s.t. } \hat{\boldsymbol{m}}=\boldsymbol{m}}, \tag{III.6}$$

$$\text{FN}_{NER} = \sum_{\boldsymbol{m}} \mathbb{1}_{\neg\exists \hat{\boldsymbol{m}} \text{ s.t. } \hat{\boldsymbol{m}}=\boldsymbol{m}}, \tag{III.7}$$

with $\text{start}(\boldsymbol{m})$ (resp. $\text{end}(\boldsymbol{m})$) the index in $\boldsymbol{d}$ of the first (resp. last) token of $\boldsymbol{m}$. By convention, this formulation has no true negatives (TN)[20]. We use the micro aggregation to compute the F1 score, precision (P), and recall (R). As a side note, F1 micro equals the accuracy because we are in a single-label prediction setting. We have:

$$\text{P}_{NER} = \frac{\text{TP}_{NER}}{\text{TP}_{NER} + \text{FP}_{NER}}, \tag{III.8}$$

$$\text{R}_{NER} = \frac{\text{TP}_{NER}}{\text{TP}_{NER} + \text{FN}_{NER}}, \tag{III.9}$$

$$\text{F1}_{NER} = \frac{2\,\text{P}_{NER}\,\text{R}_{NER}}{\text{P}_{NER} + \text{R}_{NER}}. \tag{III.10}$$

[20] A TN is a span that is not a true entity nor a predicted one. Given that the number of spans evolves quadratically depending on the size of the document and entities are relatively scarce, TNs would crush TPs, FPs, and FNs, leading to undiscriminative scores. Therefore, the consensus (e.g., [66, 113]) is to remove true negatives.

III.5.2 Coreference Resolution (CR B^3)

To evaluate coreferences, we use the B^3 metric [218], which is used to compare clustering assignments to true labels. This metric is recommended and widely employed to evaluate coreference resolution models [21, 219, 220]. B^3 defines precision and recall as follows:

$$\mathrm{P}_{CR} = \mathbb{E}_{\hat{\boldsymbol{m}}=\boldsymbol{m},\hat{\boldsymbol{m}}'=\boldsymbol{m}'} P(\boldsymbol{e} = \boldsymbol{e}' | \hat{\boldsymbol{e}} = \hat{\boldsymbol{e}}'), \tag{III.11}$$

$$\mathrm{R}_{CR} = \mathbb{E}_{\hat{\boldsymbol{m}}=\boldsymbol{m},\hat{\boldsymbol{m}}'=\boldsymbol{m}'} P(\hat{\boldsymbol{e}} = \hat{\boldsymbol{e}}' | \boldsymbol{e} = \boldsymbol{e}'), \tag{III.12}$$

with $\boldsymbol{m}$ and $\boldsymbol{m}'$ two distinct true mentions occurring in the same document ($\boldsymbol{m}$ belongs to $\boldsymbol{e}$, and $\boldsymbol{m}'$ belongs to $\boldsymbol{e}'$), and $\hat{\boldsymbol{m}}$ and $\hat{\boldsymbol{m}}'$ the corresponding predicted mentions ($\hat{\boldsymbol{m}}$ belongs to $\hat{\boldsymbol{e}}$, and $\hat{\boldsymbol{m}}'$ belongs to $\hat{\boldsymbol{e}}'$). If a true mention does not match any predicted mention, then a placeholder $\hat{\boldsymbol{m}}$ belonging to a specific error entity is created. Conversely, if a predicted mention does not match any true mention, then a placeholder $\boldsymbol{m}$ belonging to a specific error entity is created. Expected values for P_{CR} and R_{CR} are computed by iterating over all possible pairs of mentions. B^3 is calculated as an F1 score:

$$\mathrm{B}^3{}_{CR} = \frac{2\,\mathrm{P}_{CR}\,\mathrm{R}_{CR}}{\mathrm{P}_{CR} + \mathrm{R}_{CR}}. \tag{III.13}$$

III.5.3 Joint Named Entity Recognition & Coreference Resolution (Entity F1)

To provide a global metric to evaluate the extraction of entities (considering NER and CR), a simple approach is to define a hard equality between $\boldsymbol{e}$ and $\hat{\boldsymbol{e}}$:

$$\hat{\boldsymbol{e}} = \boldsymbol{e} \iff \hat{\mathfrak{e}} = \mathfrak{e} \wedge \{\hat{\boldsymbol{m}} \in \hat{\boldsymbol{e}}\} = \{\boldsymbol{m} \in \boldsymbol{e}\}, \tag{III.14}$$

the set equality is calculated with the mention equality as defined in equation (III.4). P^{H}_{entity}, R^{H}_{entity}, and $\mathrm{F1}^{H}_{entity}$ are derived in a similar fashion as equations (III.8) to (III.10).

However, these metrics strongly penalize coreferences mistakes. For instance, if a predicted entity corresponds exactly to a true entity, except that it is missing a single coreference, it will be counted as an error with the same negative impact as a completely missed entity. This results in metrics with a low power of discrimination between models. To solve this shortcoming, Zaporojets et al. [21] propose soft metrics instead. They redefine TP, FP, and FN so they are not integers but real numbers characterizing the similarity level between the predicted and true mentions of entities:

$$\mathrm{TP}^{\mathrm{P}}_{entity} = \sum_{\mathfrak{e}} \sum_{\hat{\boldsymbol{e}} \text{ s.t. } \hat{\mathfrak{e}}_{\hat{\boldsymbol{e}}}=\mathfrak{e}} \frac{|\{\hat{\boldsymbol{m}} \in \hat{\boldsymbol{e}}\} \cap \{\boldsymbol{m} \in \boldsymbol{e} \text{ s.t. } \mathfrak{e}_{\boldsymbol{e}} = \mathfrak{e}\}|}{|\{\hat{\boldsymbol{m}} \in \hat{\boldsymbol{e}}\}|}, \tag{III.15}$$

$$\mathrm{TP}^{\mathrm{R}}_{entity} = \sum_{\mathfrak{e}} \sum_{\boldsymbol{e} \text{ s.t. } \mathfrak{e}_{\boldsymbol{e}}=\mathfrak{e}} \frac{|\{\boldsymbol{m} \in \boldsymbol{e}\} \cap \{\hat{\boldsymbol{m}} \in \hat{\boldsymbol{e}} \text{ s.t. } \hat{\mathfrak{e}}_{\hat{\boldsymbol{e}}} = \mathfrak{e}\}|}{|\{\boldsymbol{m} \in \boldsymbol{e}\}|}, \tag{III.16}$$

$$\mathrm{FP}_{entity} = |\{\hat{\boldsymbol{e}} \in \mathcal{D}\}| - \mathrm{TP}^{\mathrm{P}}_{entity}, \tag{III.17}$$

$$\mathrm{FN}_{entity} = |\{\boldsymbol{e} \in \mathcal{D}\}| - \mathrm{TP}^{\mathrm{R}}_{entity}, \tag{III.18}$$

with the intersection between true and predicted mentions computed with the mention equality defined in equation (III.4). They define two versions of the true positives: TP^{P}_{entity} is normalized by the predicted mentions (precision orientation), and TP^{R}_{entity} by the true mentions (recall orientation). Then, Zaporojets et al. derive:

$$\mathrm{P}^{S}_{entity} = \frac{\mathrm{TP}^{\mathrm{P}}_{entity}}{\mathrm{TP}^{\mathrm{P}}_{entity} + \mathrm{FP}_{entity}}, \tag{III.19}$$

$$\mathrm{R}^{S}_{entity} = \frac{\mathrm{TP}^{\mathrm{R}}_{entity}}{\mathrm{TP}^{\mathrm{R}}_{entity} + \mathrm{FN}_{entity}}, \tag{III.20}$$

$$\mathrm{F1}^{S}_{entity} = \frac{2\,\mathrm{P}^{S}_{entity}\,\mathrm{R}^{S}_{entity}}{\mathrm{P}^{S}_{entity} + \mathrm{R}^{S}_{entity}}. \tag{III.21}$$

III.5.4 Relation Extraction (Relation F1)

Comparing entities composed of multiple mentions is not trivial, but it is even more complex for relations. Firstly, we can build hard metrics $\mathrm{P}^{H}_{relation}$, $\mathrm{R}^{H}_{relation}$, $\mathrm{F1}^{H}_{relation}$, following a principle similar to the last section. We define the relation equality as:

$$\hat{\boldsymbol{r}} = \boldsymbol{r} \iff \hat{\mathbb{r}} = \mathbb{r} \wedge \hat{\boldsymbol{e}}_{head} = \boldsymbol{e}_{head} \wedge \hat{\boldsymbol{e}}_{tail} = \boldsymbol{e}_{tail}, \tag{III.22}$$

the entity equality is computed with equation (III.14). However, it is too strict to eliminate an entity and all its relations if it is missing only one coreference. Zaporojets et al. [21] also propose a soft Relation F1 score, which tackles this problem. In brief, it compares the relations at a mention level, checking that both predicted mentions correspond to gold mentions and that there is this relation between them. Then, the results are aggregated at the entity level. They define $\mathrm{TP}^{\mathrm{P}}_{relation}$ and $\mathrm{TP}^{\mathrm{R}}_{relation}$ as follows:

$$\mathrm{TP}^{\mathrm{P}}_{relation} = \sum_{\mathbb{r}} \sum_{\hat{\boldsymbol{r}} \text{ s.t. } \hat{\mathbb{r}}_{\hat{\boldsymbol{r}}}=\mathbb{r}} \frac{\left| \begin{array}{l} \{(\hat{\boldsymbol{m}}_{head}, \hat{\boldsymbol{m}}_{tail}) \text{ s.t. } \hat{\boldsymbol{m}}_{head} \in \hat{\boldsymbol{e}}_{head}, \hat{\boldsymbol{m}}_{tail} \in \hat{\boldsymbol{e}}_{tail}\} \cap \\ \quad \{(\boldsymbol{m}_{head}, \boldsymbol{m}_{tail}) \text{ s.t. } \boldsymbol{m}_{head} \in \boldsymbol{e}_{head}, \boldsymbol{m}_{tail} \in \boldsymbol{e}_{tail}, \mathbb{r}_{\boldsymbol{r}} = \mathbb{r}\} \end{array} \right|}{|\{(\hat{\boldsymbol{m}}_{head}, \hat{\boldsymbol{m}}_{tail}) \text{ s.t. } \hat{\boldsymbol{m}}_{head} \in \hat{\boldsymbol{e}}_{head}, \hat{\boldsymbol{m}}_{tail} \in \hat{\boldsymbol{e}}_{tail}\}|}, \tag{III.23}$$

$$\mathrm{TP}^{\mathrm{R}}_{relation} = \sum_{\mathbb{r}} \sum_{\boldsymbol{r} \text{ s.t. } \mathbb{r}_{\boldsymbol{r}}=\mathbb{r}} \frac{\left| \begin{array}{l} \{(\boldsymbol{m}_{head}, \boldsymbol{m}_{tail}) \text{ s.t. } \boldsymbol{m}_{head} \in \boldsymbol{e}_{head}, \boldsymbol{m}_{tail} \in \boldsymbol{e}_{tail}\} \cap \\ \quad \{(\hat{\boldsymbol{m}}_{head}, \hat{\boldsymbol{m}}_{tail}) \text{ s.t. } \hat{\boldsymbol{m}}_{head} \in \hat{\boldsymbol{e}}_{head}, \hat{\boldsymbol{m}}_{tail} \in \hat{\boldsymbol{e}}_{tail}, \hat{\mathbb{r}}_{\hat{\boldsymbol{r}}} = \mathbb{r}\} \end{array} \right|}{|\{(\boldsymbol{m}_{head}, \boldsymbol{m}_{tail}) \text{ s.t. } \boldsymbol{m}_{head} \in \boldsymbol{e}_{head}, \boldsymbol{m}_{tail} \in \boldsymbol{e}_{tail}\}|}, \tag{III.24}$$

$$\mathrm{FP}_{relation} = |\{\hat{\boldsymbol{r}} \in \mathcal{D}\}| - \mathrm{TP}^{\mathrm{P}}_{relation}, \tag{III.25}$$

$$\mathrm{FN}_{relation} = |\{\boldsymbol{r} \in \mathcal{D}\}| - \mathrm{TP}^{\mathrm{R}}_{relation}, \tag{III.26}$$

with the intersection between the pairs of predicted and true mentions computed with equation (III.4) ($\hat{\boldsymbol{m}}_{head} = \boldsymbol{m}_{head} \wedge \hat{\boldsymbol{m}}_{tail} = \boldsymbol{m}_{tail}$). $P^S_{relation}$, $R^S_{relation}$, and $F1^S_{relation}$ are calculated as defined in equations (III.19) to (III.21).

III.5.5 Entity Linking (Hit@1, Hit@5, NF, MR)

To evaluate entity linking, we propose to use the Hit@1, Hit@5, Not Found, and Mean Rank metrics:

- Hit@1. The proportion of entities where the correct resource is the first candidate returned by the entity linker.
- Hit@5. The proportion of entities where the correct resource is in the first five candidates returned by the entity linker.
- Not Found (NF). The proportion of entities where the entity linker does not find the correct resource.
- Mean Rank (MR). The average rank where the correct resource is found (only for entities for which one candidate is correct).

We have the same aggregation problem for these metrics, as our predicted entities are not strictly equal to the gold entities. We employ the same idea as Entity F1 by comparing entities at the mention level and then aggregating the comparisons at the entity level. We suppose that each predicted entity $\hat{\boldsymbol{e}}$ is associated with an ordered list of candidate resources for the entity linking: $\hat{\boldsymbol{o}}(\hat{\boldsymbol{e}}) = [\hat{o}_0, \hat{o}_1, ...]$, $\hat{o}_0$ being the most related candidate. We give each candidate $\hat{o}$ as score f_{EL}, corresponding to its index in $\hat{\boldsymbol{o}}(\hat{\boldsymbol{e}})$:

$$f_{EL}(\hat{o}_i|\hat{\boldsymbol{e}}) = \begin{cases} i & \text{if } \hat{o}_i \in \hat{\boldsymbol{o}}(\hat{\boldsymbol{e}}), \\ |\hat{\boldsymbol{o}}(\hat{\boldsymbol{e}})| + 1 & \text{otherwise.} \end{cases} \tag{III.27}$$

As a side note, if the entity linking model returns probabilities associated with each $\hat{o}$, they can be used in place of $f_{EL}(\hat{o}_i|\hat{\boldsymbol{e}})$. Then, for each mention of a gold entity, we find the corresponding predicted mention, if it exists, using equation (III.4). With this, we get the ordered list of candidates associated with the mention. To aggregate the candidates for all the mentions of a gold entity, we sum the $f_{EL}(\hat{o}_i|\hat{\boldsymbol{e}})$:

$$f_{EL}(\boldsymbol{e}, \hat{o}_i) = \sum_{\hat{\boldsymbol{e}}, \hat{\boldsymbol{m}} \in \hat{\boldsymbol{e}}, \boldsymbol{m} \in \boldsymbol{e} \text{ s.t. } \hat{\boldsymbol{m}} = \boldsymbol{m}} f_{EL}(\hat{o}_i|\hat{\boldsymbol{e}}). \tag{III.28}$$

The ranking is obtained by sorting the scores in ascending order, with the candidate with the lower score being the best. Hit@1, Hit@5, NF, and MR are calculated with this ranking. As an aside, NUM or TIME entities are ignored during this evaluation, as they are not disambiguated.

III.5.6 Unsupervised and Open-World Metrics

The metrics defined in the previous sections are adapted to evaluate supervised, few-shot, or zero-shot baselines. They need, however, minor modifications to evaluate unsupervised or open-world baselines where the set of true entity/relation types is not equal to the set of predicted entity/relation types. This challenge is significant for unsupervised models: as they do not know the target classes, the predicted clusters cannot be directly mapped to the true entity types (no direct link between cluster IDs and class IDs). This problem impacts NER (entity types) and RE (relation types). As a result, traditional classification metrics such as precision, recall, and F1 score cannot be used to evaluate unsupervised IE models.

Multiple metrics have been proposed to compare clustering to true labels. Compared to classification metrics, they are robust to permutations (meaning the cluster IDs will not impact the final score) and to partial matches (e.g., two or more clusters that correspond to a single class or the opposite). We implement four widely used metrics: B^3 [218], V-measure [221], Adjusted Mutual Information (AMI) [2, 222], and Adjusted Rand Index (ARI) [3, 4].

B^3 and V-measure provide alternative definitions for precision (homogeneity for V-measure) and recall (completeness for V-measure) and allow the calculation of an F1 score. V-measure tends to penalize more small impurities in a pure cluster than in a less pure cluster, where B^3 has a more linear behavior [223].

AMI and ARI give a single value. The main advantage of AMI and ARI over V-measure and B^3 is that they are adjusted for chance: a random clustering will reliably produce AMI and ARI close to 0. Additionally, they are defined over $[-1, 1]$: scores below zero mean methods that are less effective than random clustering. Romano et al. [224] recommends using AMI with datasets having unbalanced class distributions and ARI with balanced distributions.

To apply these metrics for NER or RE, the correspondences between the true mentions (resp. relations) and predicted mentions (resp. relations) is calculated with the equalities defined in equations (III.4) and (III.22)[21]. A "predicted" placeholder is created with a specific error entity (resp. relation) type for the true mentions (resp. relations) that were not predicted (FN). Conversely, for the predicted mentions (resp. relations) that do not exist in the ground truth (FP), a "true" placeholder with a specific error entity (resp. relation) type is created.

We recall the definition of the four metrics in the following paragraphs. We take the example of NER. $\boldsymbol{m}$, $\hat{\boldsymbol{m}}$, $\mathfrak{e}$, or $\hat{\mathfrak{e}}$ are considered to be random variables in the definitions. To derive the actual values, the reader has to enumerate all $\boldsymbol{m} \in \mathcal{D}$ and $\hat{\boldsymbol{m}} \in \mathcal{D}$.

B^3 Similarly to section III.5.2, we have:

$$\mathrm{P} = \mathbb{E}_{\hat{\boldsymbol{m}}=\boldsymbol{m},\hat{\boldsymbol{m}}'=\boldsymbol{m}'} P(\mathfrak{e} = \mathfrak{e}' | \hat{\mathfrak{e}} = \hat{\mathfrak{e}}'), \tag{III.29}$$

$$\mathrm{R} = \mathbb{E}_{\hat{\boldsymbol{m}}=\boldsymbol{m},\hat{\boldsymbol{m}}'=\boldsymbol{m}'} P(\hat{\mathfrak{e}} = \hat{\mathfrak{e}}' | \mathfrak{e} = \mathfrak{e}'), \tag{III.30}$$

$$\mathrm{B}^3 = \frac{2\,\mathrm{P}\,\mathrm{R}}{\mathrm{P}+\mathrm{R}}, \tag{III.31}$$

[21] The types checks $\hat{\mathfrak{e}} = \mathfrak{e}$ or $\hat{\mathfrak{r}} = \mathfrak{r}$ are disabled (as there is no direct mapping between types and clusters). These checks are directly integrated into B^3, V-measure, ARI, or AMI.

V-measure V-measure defines homogeneity and completeness, relying on the conditional Shannon entropy [2]:

$$\text{Homogeneity} = 1 - \frac{\mathrm{H}(\hat{\mathfrak{e}}|\mathfrak{e})}{\mathrm{H}(\hat{\mathfrak{e}})}, \tag{III.32}$$

$$\text{Completeness} = 1 - \frac{\mathrm{H}(\mathfrak{e}|\hat{\mathfrak{e}})}{\mathrm{H}(\mathfrak{e})}, \tag{III.33}$$

$$\mathrm{V} = \frac{2\,\text{Homogeneity} \cdot \text{Completeness}}{\text{Homogeneity} + \text{Completeness}}, \tag{III.34}$$

with H the Shannon entropy. Homogeneity has a similar interpretation as precision, and completeness as recall.

Adjusted Rand Index The Rand Index is defined as the probability that the clustering and true labels assignment are compatible:

$$\mathrm{RI} = \mathbb{E}_{\hat{\boldsymbol{m}}=\boldsymbol{m},\hat{\boldsymbol{m}}'=\boldsymbol{m}'} P(\hat{\mathfrak{e}} = \hat{\mathfrak{e}}' \Leftrightarrow \mathfrak{e} = \mathfrak{e}). \tag{III.35}$$

The Adjusted Rand Index is the adjustment for chance of the RI, such that a random clustering will produce scores close to or equal to zero.

Adjusted Mutual Information The Mutual Information score measures the mutual dependence between the true entity type and predicted entity type random variables. It is defined as:

$$\begin{aligned} \mathrm{MI} &= \mathrm{H}(\mathfrak{e}) - \mathrm{H}(\mathfrak{e}|\hat{\mathfrak{e}}), \\ &= \mathrm{H}(\hat{\mathfrak{e}}) - \mathrm{H}(\hat{\mathfrak{e}}|\mathfrak{e}). \end{aligned} \tag{III.36}$$

Similarly to ARI, the Adjusted Mutual Information is the adjustment for chance of the MI. In practice, AMI is very related to the V-measure, as the V-measure corresponds to the Normalized Mutual Information, another method to normalize the MI. V-measure is not adjusted for chance.

III.6 Experiments

III.6.1 Baseline

As seen in section III.2, an end-to-end IE model can be seen as a four-module process with NER, CR, EL, and RE. Our objective for this baseline is to provide a simple IE model with results comparable to those of current state-of-the-art approaches. Recent papers that use DocRED as a benchmark focus on document-level RE [176, 178, 180, 186], ignoring NER, coreference resolution, and entity linking. Additionally, no document-level end-to-end IE model exists (see chapter II). Therefore, we propose implementing an IE pipeline, that is, a multi-stage model, where the four subtasks are implemented with specialized modules. As a side note, very recent models start to tackle jointly multiple tasks in a document-level setting [126, 141, 175], but

they were not available at the time Linked-DocRED was designed. The only exception is the work of Verlinden et al. [20] that jointly predicts NER, CR, and RE, which we included in our evaluation.

Named Entity Recognition We propose to use the simple yet effective span-based NER proposed by Zhong et al. [113] (PURE). This model relies on BERT [6], which can only handle documents with at most 512 tokens. As we have documents with more than 512 tokens, we propose to replace BERT with Longformer [225], which can encode documents up to 4,096 tokens, with only a marginal decrease in performance compared to BERT.

Coreference Resolution We propose implementing a well-used model, NeuralCoref[22]. This model uses NER, parsing, and pos-tagging features to predict coreferences.

Relation Extraction We do not use the DocRED baseline, as it is based on Bi-LSTMs and GloVe embeddings [89], which no longer correspond to the best state-of-the-art models, such as those based on large language models. Similarly to Prieur et al. [44], we propose to use ATLOP [178] to extract relations. Contrary to concurrent approaches (e.g., [21, 102, 176, 177, 226]), which often represent the knowledge explicitly as a graph that can be processed with Graph Neural Networks (GNN) for inference; Zhou et al. [178] propose to use implicit knowledge representations produced with BERT, which results in a simple, efficient and effective model.

In the rest of the paper, we call this NER-CR-RE ensemble **PNA** (for **P**URE [113], **N**euralCoref, and **A**TLOP [178]). This pipeline is trained using the hyperparameter values proposed by the authors of PURE [113], NeuralCoref, and ATLOP [178].

Entity Linking We propose two very simple models: *EL-Wikidata* and *EL-Wikipedia* because entity linking has not been studied much in the context of end-to-end IE models ([20, 21, 44] use very basic approaches). For EL-Wikidata, we search each mention $\hat{\boldsymbol{m}}$ of a predicted entity $\hat{\boldsymbol{e}}$ in Wikidata using the Wikidata search API. This API returns a ranked list of candidate Wikidata entities most related to the mention: $\hat{\boldsymbol{o}}(\hat{\boldsymbol{m}}) = [\hat{o}_0, \hat{o}_1, ..., \hat{o}_{|\hat{\boldsymbol{o}}(\hat{\boldsymbol{m}})|-1}]$, $\hat{o}_0$ being the best candidate. We give each candidate a score f_{EL}, corresponding to its index in $\hat{\boldsymbol{o}}(\hat{\boldsymbol{m}})$:

$$\mathrm{f}_{\mathrm{EL}}(\hat{o}_i|\hat{\boldsymbol{m}}) = \begin{cases} i & \text{if } \hat{o}_i \in \hat{\boldsymbol{o}}(\hat{\boldsymbol{m}}), \\ |\hat{\boldsymbol{o}}(\hat{\boldsymbol{m}})| + 1 & \text{otherwise.} \end{cases} \tag{III.37}$$

To aggregate the candidates for all the mentions of an entity, we sum the $\mathrm{f}_{\mathrm{EL}}(\hat{o}_i|\hat{\boldsymbol{m}})$:

$$\mathrm{f}_{\mathrm{EL}}(\hat{o}_i) = \sum_{\hat{\boldsymbol{m}} \in \hat{\boldsymbol{e}}} \mathrm{f}_{\mathrm{EL}}(\hat{o}_i|\hat{\boldsymbol{m}}). \tag{III.38}$$

The ranking is obtained by sorting the scores in ascending order, with the candidate with the lower score being the best. EL-Wikipedia follows the same principle as EL-Wikidata, replacing the Wikidata search API with Wikipedia.

[22] Available at https://github.com/huggingface/neuralcoref.

Table III.3: Evaluation of the PNA baseline and other IE models on the development split of Linked-DocRED using our proposed entity-level metrics. For Entity F1 and Relation F1, the soft metric is primarily displayed along with the hard aggregation in parenthesis. During evaluation, the ATLOP baseline (second line) can access ground truth entities and coreferences. In the second table, the first two lines access ground truth entities and coreferences, whereas PNA entity linking is based on its imperfect entity and coreference predictions.

	Mention F1 ↑	CR B^3 ↑	Entity F1 ↑		Relation F1 ↑	
Verlinden et al. [20]	-	-	-	(71.8)	-	(25.7)
ATLOP [178]	-	-	-	-	63.4	(63.4)
PNA (ours)	77.2	80.4	83.9	(82.9)	48.9	(41.1)

		Entity Linking			
		Hit@1 ↑	Hit@5 ↑	NF ↓	MR ↓
Ground truth entities	EL-Wikipedia	52.3	61.7	32.1	2.1
	EL-Wikidata	59.0	68.5	26.3	1.7
PNA (ours)	EL-Wikipedia	46.0	53.9	40.8	2.1
	EL-Wikidata	51.1	59.1	36.2	1.7

III.6.2 Results

The evaluation results are shown in table III.3. We also display the results of an integrated, multitask IE pipeline from Verlinden et al. [20] and the RE model ATLOP [178] with ground truth entities and coreferences. All methods are trained on the train split of Linked-DocRED and evaluated on its developement split. For Entity F1 and Relation F1, we show the soft and hard metrics (the hard aggregation is in parenthesis, to compare with [20, 178]).

Firstly, the Mention F1, B^3, and Entity F1 or PNA are superior to 75 % (resp. 77.2, 80.4, and 83.9), which is in the range of what is currently state-of-the-art for DocRED [20]. Compared to Verlinden et al. [20], our baseline obtains better results in hard Entity F1 (and Relation F1) while being much simpler to implement and run. A similar observation was made by Prieur et al. [44] on the DWIE dataset [21].

Our baseline performance for RE is relatively low when we look at table III.3. There is some error cascading, as the NER and the coreference resolver are imperfect. This is highlighted when we look at ATLOP [178] with ground truth entities and coreference. The difference in soft Relation F1 is 14.5 points (23 % of difference). This is a clear challenge of an IE pipeline: due to its sequential nature, performances are progressively decreasing as one module follows another. Taken separately, the performances of ATLOP are state-of-the-art, but with the full pipeline, a significant part of the performance is lost. Additionally, even with the ground truth entities and coreferences, the performances of ATLOP are far from perfect. They are linked to the complexity of handling documents, with a long context and meaningful information spread in several places (indeed, 40.7 % of the relations in Linked-DocRED need analyzing multiple sentences [1]). Full document-level relation extraction is challenging.

The final step in our evaluation is entity linking. Overall, we can see a small advantage for EL-Wikidata compared to EL-Wikipedia: +5.5 points for Hit@1 and Hit@5, −5 points for Not Found, and −0.5 for Mean Rank. We think it is linked to the fact that a Wikidata entity possesses multiple surface forms at the same time (`rdfs:label` or `rdfs:aliases` predicates), which helps during the API search. We observe an 8 point decrease in Hit@1, Hit@5, and Not Found metrics when we compare the performance of gold entities to those extracted with our baseline. In all cases, however, around one-third of entities are wrongly disambiguated (Not Found), and only 50 % – 60 % of entities are correctly disambiguated with the first match (Hit@1). The entity linking task is challenging, particularly with an imperfect entity and coreference extraction.

III.7 Conclusion

In this work, we introduce Linked-DocRED, to the best of our knowledge, the first large-scale, document-level IE dataset with manual annotations for entities, coreferences, relations, and entity linking. To do so, we develop a semi-automatic entity linking process that ensures human-quality annotations. We also propose a new entity-centric entity linking metrics and open-world and unsupervised metrics to finalize the definition of a complete benchmark for end-to-end IE model evaluation.

Experimental results of a strong IE pipeline baseline demonstrate the challenges linked to end-to-end IE:

- Cascading errors. Imperfect NER and CR negatively impact RE and EL, leading to a 14.5 % decrease in Relation F1 and 7.9 % in EL Hit@1.
- The complexity of handling documents. Even with perfect entities and coreferences, the RE and EL performances are far from perfect (63.4 % in Relation F1, and 59 % for EL Hit@1).

These results are essential to analyze our main research question. First, no one ever tested an end-to-end IE model covering the four IE tasks on a diverse, large-scale, manually labeled, and document-level dataset. In that regard, this experiment fully meets its objectives. However, when looking at the big picture, our goal is not only to implement an end-to-end IE model, but we also want it to be low-resource and open-world. Logically, these constraints will hurt the already not ideal performances (they are, in fact, insufficient to consider a production launch). Therefore, instead of pursuing the currently uncertain path toward end-to-end open-world and low-resource IE[23], we explore granular IE subtasks to bring significant improvements towards open-world and low-resource settings. This is the subject of our two following contributions: PromptORE (next chapter IV) and CITRUN (chapter V).

[23]The adjective “currently” has its importance. When confronted with this choice (early 2023), we did not see any research directions leading to dramatic improvements that would lead to acceptable models. Very recent models (e.g., AutoRE [126]) start to attain a decent level of performance (still far from deployable) for joint NER and RE in a document-level and open-world setting.

IV PromptORE

Prompt-Based Open-World and Unsupervised Relation Extraction

Unsupervised Relation Extraction (RE) aims to identify relations between entities in text without having access to labeled data during training. This setting is particularly relevant for domain-specific RE where no annotated dataset is available and for open-world RE where the types of relations are a priori unknown. Although recent approaches achieve promising results, they heavily depend on hyperparameters critical for their performances. Unfortunately, they fail to explain how to adjust these without labeled data, making them incompatible with a realistic unsupervised setting.

Therefore, to diminish the reliance on hyperparameters, we propose PromptORE, our "Prompt-based Open-World Relation Extraction" model. We adapt the prompt-tuning paradigm used in low-resource approaches to work in an unsupervised setting and use it to embed sentences expressing a relation. We then cluster these embeddings to discover candidate relation types and experiment with different strategies to automatically estimate an adequate number of clusters. To our knowledge, PromptORE is the first unsupervised RE model that does not need hyperparameter tuning.

Results on three general and specific domain datasets show that PromptORE consistently outperforms state-of-the-art models with a relative gain of more than 40 % in B^3, V-measure, and ARI. Qualitative analysis also indicates PromptORE's ability to identify semantically coherent clusters that closely resemble the actual relation types.

The source code of PromptORE is available in a public repository[1] and distributed under an open-source license.

Most of the work described in this chapter, including text, figures, and tables, was presented at CIKM'22 [33].

Contents

[1] Available at https://github.com/alteca/PromptORE.

IV.1 Introduction

In this chapter, we focus on Relation Extraction (RE), which consists of identifying the relation linking two entities in the context of a document.

RE is often seen as a supervised task [227], thus relying on datasets labeled with a predefined set of relations. However, this setting can be restrictive for some applications, especially domain-specific RE lacking annotated data or open-world RE where we do not know in advance the relations expressed in the dataset. Therefore, more flexible paradigms have been proposed, including distant supervision, few-shot learning, and unsupervised learning. Distant supervision [156, 157] annotates data automatically thanks to heuristics (such as Wikipedia hyperlinks or Wikidata predicates) or pre-trained models (especially Large Language Models). Few-shot learning [29] learns from a very small set of labeled instances. Unsupervised RE does not require a training dataset with labeled relations and assumes no prior knowledge about expected relation types.

Several recent open-world RE (OpenRE) approaches obtain interesting results on diverse datasets containing tens or hundreds of relation types [196, 197, 228]. They often try to compute a vector representation of the relation expressed in the sentence (relation embedding) and then cluster all the embeddings to identify groups of similar relations. Most of these methods rely on hyperparameters (e.g., number of epochs, regularization, early stopping, number of relation types, ...) that have a significant impact on their overall performance. However, tuning these hyperparameters most often requires access to labeled data, thus limiting the applicability of such models in a real-world unsupervised scenario.

We therefore propose **PromptORE**, our "**Prompt**-based **O**pen-World **R**elation **E**xtraction" model. Contrary to previous approaches, it relies on one hyperparameter at most: the target number k of relation types to be extracted. Our experiments show that even when no educated

guess can be made about k, an efficient estimate can easily be obtained automatically. Thus, to our knowledge, PromptORE is the first proposal for an unsupervised RE system that can operate in a fully unsupervised setting.

To achieve this, we first compute a relation embedding for each instance of a dataset that represents the relation type expressed in the instance. Contrary to previous approaches that fine-tuned Encoder-Only Language Models (EncLM) [196, 197, 198], we use the novel prompt-tuning paradigm. Prompt-tuning replaces the usual training by designing a prompt (i.e., a text inputted to EncLM) that elicits as much information as possible from the language model. Prompt-tuning is already used in few-shot RE [229, 230, 231, 232]. We propose to go further and adapt this paradigm to work in a fully unsupervised way. It has many benefits: 1. it does not involve training or fine-tuning the EncLM, thus removing a significant number of hyperparameters, 2. the proposed encoder is extremely simple, yet 3. we show that these prompt-based relation embeddings provide better results than current state-of-the-art methods. Usual clustering algorithms are then applied to group together the embeddings to discover relation types.

Let us summarize our main contributions:

- We propose PromptORE, a novel OpenRE model that minimizes the number of hyperparameters and provides straightforward ways to tune its only hyperparameter k in a completely unsupervised setting (section IV.3.2).
- We adapt the prompt-tuning paradigm to an unsupervised setting, which allows us to leverage more expressive embeddings than previous entity-pair representations [64, 196, 197] (section IV.3.1).
- We show that this model consistently outperforms previous state-of-the-art approaches on three datasets covering general and specific domains (section IV.5). We also demonstrate that the predicted clusters are semantically coherent and close to the true relation types (section IV.5.5).

IV.2 Related Work

Relation extraction aims to discover the binary relation $\mathbb{r}$ that links two entities mentioned in a text $\boldsymbol{d}$. A relation instance $\boldsymbol{r}$ is a triple $(\boldsymbol{e}_{head}, \boldsymbol{e}_{tail}, \mathbb{r})$: $\boldsymbol{e}_{head}$ is the subject entity of the relation, $\boldsymbol{e}_{tail}$ the object entity, and $\mathbb{r}$ the relation type. We draw the reader's eye to the potential ambiguity problem with the term "relation", which can refer to a relation instance $\boldsymbol{r}$ or a relation type $\mathbb{r}$. For clarity, we use the term relation instance $\boldsymbol{r}$ and relation interchangeably, and we distinguish relation type $\mathbb{r}$.

Even though RE from documents is the most general paradigm, the majority of unsupervised models focus on extraction within a single sentence, ignoring inter-sentence relations [233, 234], due to the complexity of document-level RE (as we have seen in chapter II). Recent approaches follow a conceptual two-step process [64, 104, 121, 235]:

1. Relation Embedding, which computes a vector representation of the relation instance;

2. Relation Type Classification.

In practice, methods can be integrated, that is, jointly modeling the two phases in a single model [22, 64]; or separated with two models trained separately [197, 223]. To compute relation embeddings, word embedding models are often used, such as GLoVe [89], ELMO [236], Bi-LSTM embeddings [15], or EncLM embeddings such as BERT [6, 15].

IV.2.1 Few-Shot Relation Extraction

This setting aims to learn a relation extraction model from the least amount of labeled data. Models focus on relation type classification: they suppose the existence of a relation between two entities [29], ignoring the case of sentences mentioning unrelated entities. Snell et al. [237] propose to use prototypical networks to determine a prototype for each relation type and predict the type by measuring the distance between the relation instance embedding and each prototype. Zhao et al. [198] and Ren et al. [238] propose to improve this method using transfer-learning from out-of-domain labeled datasets.

Recent few-shot methods consider the use of prompt-tuning with BERT and, more broadly, EncLM [239], as it allows for more efficient learning in low-resource setting [240]. Prompt-tuning replaces fine-tuning by designing a prompt, a piece of text containing the special `[MASK]` token, and asks an EncLM to predict the embedding of this `[MASK]` token. This embedding is then compared with a set of target tokens (that can be seen as relation type prototypes) to determine the type expressed in this sentence [229, 230, 231, 232]. Efforts are focused on optimizing the prompt $\mathcal{P}_{\mathcal{R}}$ and selecting a set of target tokens effectively representing the relation types. In particular, Jiang et al. [241] propose text-mining and paraphrasing-based methods to generate prompts.

The current major limitation with few-shot RE is the closed-world hypothesis, which stipulates that relation types must be known in advance (although recent papers have started to explore none-of-the-above prediction [206, 229]). Another shortcoming is the diversity of evaluation settings highlighted by Perez et al. [29]:

- Tuned few-shot learning [155, 242]. They tune the learning algorithm on a large annotated validation dataset from the same domain as the test set. Similarly, some papers do not explain how to adjust hyperparameters (especially learning rate, number of epochs, or training steps that are critical to control overfitting) without the use of an in-domain validation set [243, 244]. In this case, the term few-shot is questionable.

- Multi-distribution few-shot learning [237, 238]. They access labeled data from many domains (different from the target domain) to adjust hyperparameters or train the model to learn faster (meta-learning). This does not translate to languages where annotated data is scarce.

- True few-shot learning. No validation data (in-domain or out-of-domain) are used.

Only the last setting can be considered true few-shot learning and applies to real-world scenarios.

IV.2.2 Unsupervised Relation Extraction

Unsupervised RE aims to extract relations without having access to a labeled dataset during training. Approaches can be divided into two subgroups: triple extraction (that we presented in detail in section II.5) and relation typing. The main weakness of triplet-based approaches is the lack of predicate disambiguation. Indeed, as they rely on surface forms for the predicates, identifying the generic relation type or working with grammatically incorrect sentences (e.g., social network posts) is particularly challenging.

To solve this problem, Yao et al. [245] first proposed to learn a relation classifier using Latent-Dirichlet Allocation [246], a generative probabilistic model. This relation classifier allows grouping together relation instances with very different grammatical predicates. Nowadays, most of these relation typing methods rely on relation embeddings. They compute an embedding that encodes the underlying relation type, which is then used to identify groups of relations that share the same relation type. The earliest methods used syntactic and semantic features [245, 247, 248]. Elsahar et al. [249] add word embedding features based on GloVe [89], apply dimensionality reduction methods, and an agglomerative clustering model to identify clusters of relation instances. Marcheggiani et al. [248] use a fill-in-the-blank task: they mask one entity and try to predict it using a Variational Auto-Encoder (VAE) [250], proving the benefit of generating a supervision signal. This method is further improved by adding two regularization losses to limit overfitting [223] and by finding a more effective formulation of the VAE task [251]. However, Tran et al. [252] outperforms VAE approaches only using the entity types of the subject/object entities (in one-hot encoding) as their relational embedding, demonstrating the benefits of using available metadata.

Hu et al. [197] adopt another supervision signal with SelfORE: they compute pseudo labels using a k-means clustering on relation embeddings and train a BERT classifier to reproduce them. Using the fine-tuned BERT model, new relation embeddings are generated, clustered, and pseudo-labels are computed. SelfORE is iteratively trained in a self-learning loop. As an alternative, Wu et al. [196] propose to learn a distance metric representative of the relation types (using siamese neural networks [253]) to compare pairs of instances. This metric is learned on an annotated dataset and applied to unlabeled data to identify instances expressing similar relation types (the training and test datasets are from different domains). Lou et al. [228] use ranked list loss [254] as an alternative to siamese neural networks. Finally, recent unsupervised RE methods tend to use transfer-learning: learning some relation embeddings or metrics on general domain annotated datasets and adapting them to unsupervised data coming from a specific domain [196, 198, 228]. Compared to triples extraction, relation typing assumes that there is always a relation between the two entities, which can be seen as a limitation.

To allow evaluation of such OpenRE models, previous works tend to train them on labeled datasets and compare their predictions with ground truth relations using external clustering evaluation metrics such as V-measure [221], Adjusted Rand Index [3, 4] or B^3 [218].

Are Current OpenRE Models Completely Unsupervised? Similarly to the observation of Perez et al. [29] for few-shot learning, we notice that unsupervised RE methods are trained and evaluated with various settings. In particular, they also suffer from the hyperparameter tuning

problem approaches. Indeed, they rely extensively on hyperparameters that need to be adjusted: number of epochs or training steps [196, 197, 228, 248, 251, 252], learning rate, regularization [223], entity types [252], early-stopping [252], and most importantly the number of relations types k the model is supposed to extract [196, 197, 223, 228, 248, 249, 251, 252]. In a realistic unsupervised setting, these hyperparameters are extremely hard to determine, and cited papers do not present satisfactory methods to estimate them without labeled data. We conclude that these approaches are not fully unsupervised regarding hyperparameter tuning, which, in our opinion, restricts their use in real-world applications.

Therefore, in this chapter, we pursue the objective of exploring true unsupervised RE that does not need annotated data during its training procedure or for hyperparameter tuning. When this contribution was made (before GPT-3.5 and ChatGPT [30]), we were the first to diminish the reliance on hyperparameters to attain more unsupervised RE.

IV.3 Description of PromptORE

PromptORE aims to extract the binary relation $\boldsymbol{r}$ between two already known entities $\boldsymbol{e}_{head}$ and $\boldsymbol{e}_{tail}$ present in the same sentence. As in previous unsupervised RE works, we focus on sentence RE, even though some relations may be missed. More precisely, as we follow an unsupervised setting, the first objective of PromptORE is to group relation instances expressing the same relation type $\mathbb{r}$ without having access to labeled data during training and hyperparameter tuning. Our second objective is to minimize the number of hyperparameters that PromptORE needs and provide unambiguous procedures to adjust them without annotated data.

To achieve these goals, we suppose we have access to a dataset $\mathcal{D}$ (see figure IV.1) containing instances with the following properties:

1. An instance is described with a triple $(\boldsymbol{e}_{head}, \boldsymbol{e}_{tail}, \boldsymbol{d})$, where $\boldsymbol{d}$ is the instance text $\boldsymbol{e}_{head}$ and $\boldsymbol{e}_{tail}$ are two entities in $\boldsymbol{d}$.

2. We suppose that $\boldsymbol{e}_{head}$ and $\boldsymbol{e}_{tail}$ have already been extracted (but not typed).

3. We suppose $\boldsymbol{e}_{head}$ and $\boldsymbol{e}_{tail}$ have a single mention $\boldsymbol{m}_{head}$ and $\boldsymbol{m}_{tail}$ in $\boldsymbol{d}$[2]. To simplify mathematical notations, in this chapter only, the text of $\boldsymbol{e}_{head}$ (resp. $\boldsymbol{e}_{tail}$) is the text of its unique mention $\boldsymbol{m}_{head}$ (resp. $\boldsymbol{m}_{tail}$).

4. In the instance text $\boldsymbol{d}$, $\boldsymbol{e}_{head}$ and $\boldsymbol{e}_{tail}$ are linked by a binary relation of type $\mathbb{r}$. We do not consider the case where there is no relation between $\boldsymbol{e}_{head}$ and $\boldsymbol{e}_{tail}$.

5. We cannot access any relation label during training and hyperparameter tuning.

As a side note, properties 1–4 are standard for unsupervised RE [197, 223, 252]. $\mathcal{R}$ is the set of the k relation types in $\mathcal{D}$. We have no information about the types in $\mathcal{R}$ (no label, description, entity type compatibility, etc.). Regarding k, it can either be given by the user or automatically

[2] As an aside, this constraint is more related to the evaluation datasets (which contain only one mention per entity) than PromptORE.

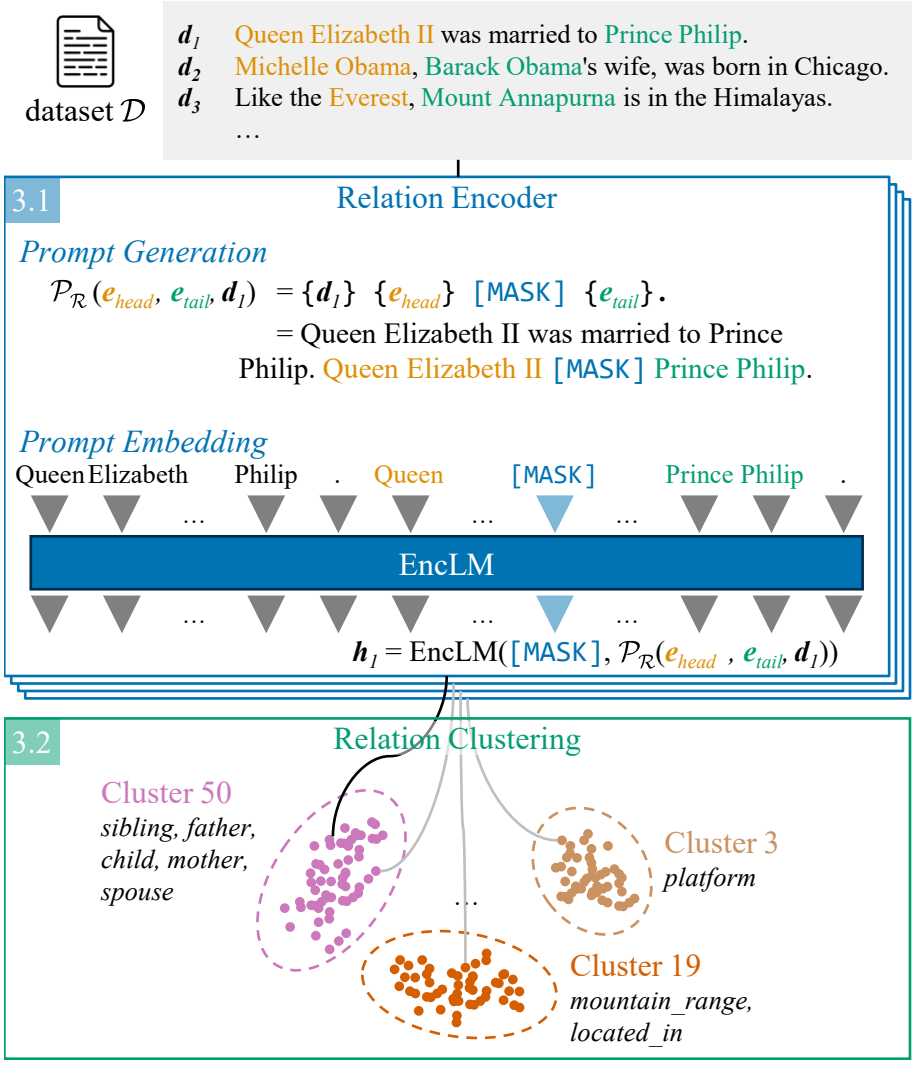


Figure IV.1: Overview of PromptORE.

estimated by methods described in section IV.3.2. As shown in figure IV.1, PromptORE is composed of two main modules, similarly to [197, 198]:

1. Relation Encoder. It computes a vector representation of the relation expressed in the current instance.

2. Relation Clustering. It clusters the relation embeddings of the whole dataset to identify groups of relation instances expected to express the same relation type $\mathbb{r}$.

IV.3.1 Relation Encoder

This module aims to compute a vector representation (or relation embedding) of the relation expressed between $\boldsymbol{e}_{head}$ and $\boldsymbol{e}_{tail}$ in the current sentence $\boldsymbol{d}$. We want this relation embedding to be representative of the underlying relation type: if the relation embeddings of two instances are close (relative to a specific distance metric), these instances convey, most probably, the same relation type. In other words, the relation encoder aims to abstract the notion of relation instance to provide embeddings that are easier to compare.

In recent papers [22, 113, 197, 198, 228, 251, 255], relation embeddings are computed using EncLM such as BERT [6, 15] or RoBERTa [256]. BERT (and other EncLMs) takes some tokenized text as input and computes an embedding for each input token, which is representative of the token itself and its context of use. BERT also includes a Masked Language-Model (MLM) head, which allows it to predict the most probable tokens associated with an embedding. In addition to real word tokens (that can be a word, subword, or punctuation), BERT uses some special virtual tokens:

- `[CLS]` and `[SEP]`. By convention, `[CLS]` needs to be inserted at the start of the text, and `[SEP]` indicates the end of the text.

- `[MASK]`. It represents a token that is hidden/unknown, and BERT will try to compute a satisfactory embedding. Then, using the MLM head, BERT can predict the most probable tokens. Thanks to this `[MASK]` token, BERT can auto-complete sentences.

We define $\boldsymbol{h} = \text{EncLM}(\texttt{[MASK]}, \mathcal{P})$ as the EncLM embedding of the `[MASK]` token in the context of $\mathcal{P}$. $\mathcal{P}$ is a text that must contain one `[MASK]` token.

Previous works [22, 113, 197, 198, 228, 251, 255] use an entity-pair representation paradigm to compute relation embeddings. A vector representation for each entity is computed (using special markups or virtual tokens, aggregation, attention, ...), and the two representations are aggregated (concatenated or merged). Computing these entity-pair representations requires a specially fine-tuned EncLM. Indeed, the concatenation of the raw EncLM embeddings of the two entities is not guaranteed to produce embeddings characteristic of the underlying relation type. This property comes after fine-tuning the EncLM specifically for that task (either in a fully supervised setting or with secondary supervision signals [196, 197]). We believe that most of the hyperparameter-tuning problem comes specifically from this step. Therefore, we opt for another method to generate relation embeddings: prompt-tuning.

Prompt-Tuning The idea behind prompt-tuning is to benefit from EncLM's abilities to predict masked tokens. It is already used in few-shot RE by [229, 230, 231, 232]. It can be summarized as follows:

1. Design a prompt $\mathcal{P}_\mathcal{R}$, which is a sequence of tokens that includes one `[MASK]` token. For instance, Lv et al. [229] use the template $\mathcal{P}_\mathcal{R}(\boldsymbol{e}_{head}, \boldsymbol{e}_{tail}, \boldsymbol{d})$ = "`{`$\boldsymbol{d}$`} In this sentence, {`$\boldsymbol{e}_{head}$`} is the [MASK] of {`$\boldsymbol{e}_{tail}$`}.`".
2. Identify a set of label tokens $\mathcal{T}$ that represents each relation type. Label tokens are generally real words chosen because of their semantic similarity with the relation type [232]. They can also be learned virtual tokens [230, 231]. In that case, they are very close to prototypes in prototypical networks.
3. Predict the `[MASK]` embedding in the context of $\mathcal{P}_\mathcal{R}$ using the EncLM:

$$\boldsymbol{h} = \text{EncLM}(\texttt{[MASK]}, \mathcal{P}_\mathcal{R}). \tag{IV.1}$$

4. With this embedding, compute the probability to predict each label token $t \in \mathcal{T}$ thanks to the MLM head of the EncLM.
5. Select the relation type represented by the label token t with the highest probability.

In itself, prompt-tuning does not require fine-tuning the EncLM but necessitates designing an optimal prompt $\mathcal{P}_\mathcal{R}$ and a set of label tokens $\mathcal{T}$. The two are usually adjusted using a labeled dataset and cannot be applied directly to our unsupervised relation encoder.

Unsupervised Prompt-based Relation Encoder For our relation encoder, we propose simplifying the general principles of prompt-tuning to work unsupervised. In practice, our solution is also significantly related to prompting (which has gained significant appeal with the rise of LLMs [257]). We propose removing the set of label tokens $\mathcal{T}$ and using the simplest prompt $\mathcal{P}_\mathcal{R}$ possible. Our proposed prompt template is as follows:

$$\mathcal{P}_\mathcal{R}(\boldsymbol{e}_{head}, \boldsymbol{e}_{tail}, \boldsymbol{d}) = \text{"`{`}\boldsymbol{d}\texttt{`} {`}}\boldsymbol{e}_{head}\texttt{`} [MASK] {`}}\boldsymbol{e}_{tail}\texttt{`}.`" \tag{IV.2}$$

with $\{\cdot\}$ variable substitution ($\{\boldsymbol{d}\}$ is replaced by the text of $\boldsymbol{d}$), $\boldsymbol{d}$ the instance text, $\boldsymbol{e}_{head}$ (resp. $\boldsymbol{e}_{tail}$) the text of the head (resp. tail) entity. We have removed the `[MASK]` and `[SEP]` tokens from our equations for simplification purposes, but they are, of course, inputted to BERT at their correct locations. For instance:

$\boldsymbol{d}$ = "Queen Elizabeth II was married to Prince Philip.",
$\boldsymbol{e}_{head}$ = "Queen Elizabeth II",
$\boldsymbol{e}_{tail}$ = "Prince Philip",
$\mathcal{P}_\mathcal{R}(\boldsymbol{e}_{head}, \boldsymbol{e}_{tail}, \boldsymbol{d})$ = "Queen Elizabeth II was married to Prince Philip. Queen Elizabeth II `[MASK]` Prince Philip.".

As we remove $\mathcal{T}$, we use the `[MASK]` embedding as our relation embedding. Thus, the relation encoder process is the following (as shown in figure IV.1):

1. Apply the template defined in equation (IV.2) to generate a prompt for the current instance.
2. Predict the embedding of the `[MASK]` token with the EncLM and use it as our relation embedding:

$$\mathrm{f_{RE}}(\boldsymbol{e}_{head}, \boldsymbol{e}_{tail}, \boldsymbol{d}) = \mathrm{EncLM}(\texttt{[MASK]}, \mathcal{P}_{\mathcal{R}}(\boldsymbol{e}_{head}, \boldsymbol{e}_{tail}, \boldsymbol{d})). \tag{IV.3}$$

Alternative Prompts If we analyze $\mathcal{P}_{\mathcal{R}}$, we can see that the EncLM will likely fill the `[MASK]` token with a verb, as it is trained to produce grammatically correct sentences. It raises questions: can all relations be expressed with a verb and a single word? We did an analysis on the 9,892 Wikidata relation types with surface forms:

- More than 75 % of relation types need two words or more to be expressed. For instance, acceptable surface forms for *birthplace* are "born in" or "the birthplace of".
- Surface forms usually contain a root word (noun, verb) and tool words. 92 % of the root words are nouns[3], and only 6.7 % verbs. The most common tool words are of, in, by, the, a, to.

Therefore, it is interesting to consider alternative prompts encouraging the EncLM to predict a noun (Lv et al. [229] also focuses on noun prediction). In addition, Lv et al. [229] introduce a prefix to their prompt: "In this sentence". Therefore, we define alternative prompt templates aiming at predicting nouns and with various prefixes:

$\mathcal{P}'^{1}_{\mathcal{R}}(\boldsymbol{d}, \boldsymbol{e}_{head}, \boldsymbol{e}_{tail})$ = "{$\boldsymbol{d}$} {$\boldsymbol{e}_{head}$} `is the [MASK] of` {$\boldsymbol{e}_{tail}$}."
$\mathcal{P}'^{2}_{\mathcal{R}}(\boldsymbol{d}, \boldsymbol{e}_{head}, \boldsymbol{e}_{tail})$ = "{$\boldsymbol{d}$} `In this sentence,` {$\boldsymbol{e}_{head}$} `is the [MASK] of` {$\boldsymbol{e}_{tail}$}."
$\mathcal{P}'^{3}_{\mathcal{R}}(\boldsymbol{d}, \boldsymbol{e}_{head}, \boldsymbol{e}_{tail})$ = "{$\boldsymbol{d}$} `We deduce that` {$\boldsymbol{e}_{head}$} `is the [MASK] of` {$\boldsymbol{e}_{tail}$}."

$\mathcal{P}'^{2}_{\mathcal{R}}$ is the same as Lv et al. [229]. In our true unsupervised setting, we cannot choose the optimal prompt from the previous ones nor use automatic methods to generate prompts (such as [241]) as they require access to labeled data. Therefore, the main results of PromptORE are computed using $\mathcal{P}_{\mathcal{R}}$, the most straightforward prompt of all. In a secondary phase, we will analyze PromptORE's performances with these alternative prompts.

IV.3.2 Relation Clustering

Similarly to previous unsupervised and few-shot RE models, we measure the similarity between two EncLM embeddings using the Euclidian distance[4] [197, 198, 249]. We then cluster the relation embeddings computed on the entire dataset $\mathcal{D}$ to find groups of relation instances that are close, and we expect these clusters to be suitable candidate relation types.

[3] In practice, they are often used with the verb "be" (e.g., "is the birthplace of").
[4] Strictly speaking, the cosine similarity is generally preferred to compare embeddings, but clustering algorithms are often designed with the Euclidian distance in mind.

K-Means Clustering If we know the number of relation types k in advance, we propose using a simple k-means clustering [258, 259].

Clustering without k The most general case, however, is that we do not know k. To tackle this problem, we can take two points of view: use clustering models that do not require a predefined number of clusters or estimate k automatically and use it with regular clustering algorithms.

For the first point of view, multiple models are available, the main ones being Agglomerative Clustering (HAC) [260], DBSCAN [261], OPTICS [262], Infinite Gaussian Mixture Model (IGMM) [263], or Affinity Propagation [264]. Nevertheless, most cannot be applied in our case: HAC, DBSCAN, or IGMM need other hyperparameters (such as density), and Affinity Propagation does not scale well to big datasets. Therefore, we propose to use OPTICS.

As a second alternative, we propose implementing the elbow rule [265] to estimate the number of clusters. The idea of the elbow rule is to select an upper bound K to the number of clusters and compute a clustering (e.g., k-means clustering) for each $k, 2 \leq k \leq K$. For each clustering, we measure its quality using an internal metric (not relying on external data, such as labels we do not have) that analyzes the geometric structure of the proposed partitioning. Widely used internal metrics for the elbow rule are the silhouette coefficient [266], the distortion score (sum of the squared distance between the points and their cluster centroids), or the Calinski-Harabasz score [267]. Intuitively, we expect that increasing the number of clusters will improve the value of these internal metrics since there are more parameters to explain the data. However, when we have more clusters than the actual number of relation types, these scores will most likely grow more slowly as we can only subdivide actual relation types. We thus anticipate an inflection point in the internal metric curve. The elbow rule aims to find this inflection point, the "elbow" in the curve: the optimal trade-off between a reasonable number of clusters and a high silhouette coefficient. The elbow location is generally found visually, but automatic methods are also available [268, 269], although less precise than the human eye.

IV.4 Experimental Setup

IV.4.1 Datasets

To evaluate PromptORE, we exploit labeled datasets, which are only used during evaluation. As we have said in section III.3, we are only interested in sentence-level RE, which is already very challenging for unsupervised models. It explains why we omitted the Linked-DocRED dataset we created in chapter III.

The first dataset we choose is FewRel [210][5]. This dataset comprises text from Wikipedia pages that have been automatically annotated by aligning the text with Wikidata triples (distant-supervision setting), then manually checked for each instance. As we evaluate unsupervised RE, there is no need for training data, so we merge the training and development split together and evaluate PromptORE and the baselines on them. FewRel contains 80 relation types, with 700

[5] Available at https://github.com/thunlp/FewRel.

instances each (20 other relations are available in the test set, which is kept private). Therefore, the dataset is composed of 56,000 instances. Finally, FewRel is not multilabel: it contains, at most, one relation for each pair of entities.

FewRel is a general domain dataset, but we also want to evaluate PromptORE on more specific domains. Our second dataset is FewRel NYT [206]. This time, the sentences are taken from newspaper articles in the New York Times. It is also automatically annotated and manually checked using Wikidata. FewRel NYT contains 25 different relation types, with 100 instances each.

Our third dataset is FewRel PubMed [206]. Texts come from PubMed, a database of biomedical literature. It is also automatically annotated (this time with the UMLS knowledge base) and manually checked. It comprises 10 relation types, with 100 instances each.

IV.4.2 Metrics

Traditional classification metrics such as accuracy, precision, recall, or F1 score cannot be used to evaluate and compare PromptORE's performances. There is no direct link between our cluster IDs and the relation types. Therefore, we use the metrics presented in section III.5.6, namely, B^3, V-measure, and the Adjusted Rand Index (ARI). We choose the ARI over the Adjusted Mutual Information because the evaluated datasets have balanced class distributions. These three metrics take complementary points of view: ARI is based on pairwise similarity (enumerating all pairs of instances), B^3 on one instance versus the dataset, and the V-measure on clusters.

To analyze the experimental results precisely, let us recall some metrics properties. If a model always predicts the most frequent class, V-measure, and ARI will equal 0. If a model predicts a random distribution, ARI will be close to 0 because ARI is adjusted for chance.

IV.4.3 Baselines

We compare PromptORE with the state-of-the-art approach *SelfORE* [197], and two previous approaches based on Variational Auto-Encoders, *EType+* [252] and *UIE-PCNN* [223]).

SelfORE encodes instances with BERT and clusters these embeddings with an adaptive clustering method to generate pseudo labels that are finally used to train a classifier. It is trained self-supervised: the classifier generates new pseudo-labels to improve BERT embeddings.

UIE-PCNN encodes instances with a Piecewise Convolutional Neural Network (PCNN) [270] and uses a Variational Auto-Encoder to classify instances in an unsupervised way. Additionally, they propose two regularization losses, skewness and dispersion, to fight against the VAE's tendency to predict a single relation or a uniform distribution. Since UIE-PCNN relies on PCNN, an older embedding method, we propose to replace it with a BERT model (similarly to [252]), and we call this method *UIE-BERT*.

EType+ shows that using only entity types to encode instances provides better results than UIE-PCNN. They propose a simple typing schema for the entities: Organization, Person, Location, Miscellaneous. We expect the performances to be lower on domain-specific datasets (FewRel NYT and FewRel PubMed) with more specific entity types.

IV.4.4 Implementation Details

For PromptORE, we suppose we do not know k, except in section IV.5.1. Besides, we use the *bert-base-uncased* weights to initialize BERT. We also implement a model with RoBERTa embeddings (using the *roberta-base* pre-trained parameter). We use the scikit-learn implementation of OPTICS and k-means[6]. There are no hyperparameters to adjust.

For SelfORE, we use their publicly available implementation; for EType+ and UIE-PCNN, we use the implementation of Tran et al. [252]. The baselines are trained knowing the correct number of relations k (i.e., 80 for FewRel, 25 for FewRel NYT, and 10 for FewRel PubMed). All baselines are trained with the hyperparameter values determined by their authors.

IV.5 Results & Analysis

IV.5.1 Comparison With the Baselines

In this section only, to allow a fairer comparison with previous approaches, PromptORE knows k, the number of different relations, and a k-means clustering is used. Table IV.1 shows the results of the models on our three datasets. An informed reader may find the reported results of SelfORE evaluated on FewRel in table IV.1 low compared to those shown in other publications [51, 198, 201, 202, 271] (with an F1 B^3 between 25 % – 30 % instead of 45 % – 55 %). However, in these cases, SelfORE was evaluated on the development set of FewRel with only 16 relation types and 11,200 instances [51, 201, 202, 228]; or a subset of 1,600 instances with 16 relation types [198, 271]. We could reproduce their results using the same sampling procedure. We do not use these settings in our evaluation, as it is an easier task than FewRel with its 80 different relation types.

PromptORE consistently outperforms SelfORE, the previous state-of-the-art method, with a gap of 19 % in B^3, 18 % in V-measure and 19 % in ARI on FewRel. It represents a relative gain in performance of more than 40 %. The performance gap is even more significant with UIE-BERT, UIE-PCNN, and EType+. We observe similar conclusions with the two other datasets.

Looking more closely, we notice that UIE-BERT [223] obtains abysmal results on all three datasets: it always predicts the same relation. Strangely, Simon et al. [223] proposed two regularization losses to avoid precisely this single-class prediction situation, but this problem with UIE-PCNN and UIE-BERT was also observed by Yuan et al. [251] and Tran et al. [252]. We believe this is due to the hyperparameter that controls the balance between classification and regularization losses, which needs to be explicitly fine-tuned for each dataset. This model is symptomatic of the hyperparameter tuning problem of previously state-of-the-art unsupervised RE: performances plummet when labeled validation data is unavailable.

Finally, we notice a very small difference in performance between BERT and RoBERTa embeddings with PromptORE. In practice, both PLMs are well suited to provide precise results, and we decide to use BERT embeddings for the following parts of this paper.

[6]Available at https://scikit-learn.org.

Table IV.1: RE performances (in %) of PromptORE and previous state-of-the-art baselines on three datasets. PromptORE (and the other baselines) know the number of relations k. The best B^3 F1, V-measure F1, and ARI for each dataset is in **bold**.

Model	B^3			V-measure			ARI
	P	R	F1	Hom.	Comp.	F1	
FewRel							
UIE-PCNN	5.2	6.8	5.9	21.1	21.6	21.3	4.9
UIE-BERT	1.3	100	2.5	0	100	0	0
EType+	7.5	8.0	13.7	33.3	79.1	47.9	8.4
SelfORE	24.4	36.3	29.2	50.4	56.6	53.2	24.4
PromptORE (RoBERTa)	47.8	47.9	47.9	71.2	72.5	**71.8**	**43.7**
PromptORE (BERT)	48.7	48.8	**48.8**	71.0	72.7	**71.8**	43.4
FewRel NYT							
UIE-PCNN	7.3	27.1	11.5	9.6	15.8	11.9	3.0
UIE-BERT	4.0	100	7.8	0	100	0	0
EType+	11.0	92.6	19.6	23.0	84.9	36.2	7.8
SelfORE	32.4	48.1	38.7	50.0	58.9	54.1	26.8
PromptORE (RoBERTa)	62.6	65.3	63.9	75.7	78.1	76.8	**57.3**
PromptORE (BERT)	63.7	66.6	**65.1**	76.5	79.5	**78.0**	56.9
FewRel PubMed							
UIE-PCNN	14.4	45.2	21.9	10.3	19.2	13.5	7.2
UIE-BERT	10.0	100	18.2	0	100	0	0
EType+	10.0	100	18.1	0	100	0	0
SelfORE	53.7	66.1	59.3	58.8	68.7	63.4	45.4
PromptORE (RoBERTa)	73.7	73.2	73.5	76.5	77.2	76.9	68.1
PromptORE (BERT)	77.6	77.2	**77.4**	81.0	81.2	**81.1**	**73.8**

Table IV.2: Comparison of the RE performances (in %) of PromptORE with different prompts on three datasets. PromptORE is trained with the exact number of relations k. The best B^3 F1, V-measure F1, and ARI for each dataset is in **bold**.

Dataset	Prompt	B^3 F1	V-measure F1	ARI
FewRel	$\mathcal{P}_\mathcal{R}$ (PromptORE)	48.8	71.8	43.4
	$\mathcal{P}_\mathcal{R}^\emptyset$	33.8	57.4	28.8
	$\mathcal{P}_\mathcal{R}^{'1}$	48.9	71.7	44.5
	$\mathcal{P}_\mathcal{R}^{'2}$	49.4	72.4	46.3
	$\mathcal{P}_\mathcal{R}^{'3}$	**50.5**	**73.0**	**47.7**
FewRel NYT	$\mathcal{P}_\mathcal{R}$ (PromptORE)	65.1	**78.0**	56.9
	$\mathcal{P}_\mathcal{R}^\emptyset$	51.3	65.7	41.6
	$\mathcal{P}_\mathcal{R}^{'1}$	**65.8**	77.8	**62.0**
	$\mathcal{P}_\mathcal{R}^{'2}$	61.0	74.8	56.9
	$\mathcal{P}_\mathcal{R}^{'3}$	65.6	77.7	61.7
FewRel PubMed	$\mathcal{P}_\mathcal{R}$ (PromptORE)	77.4	**81.1**	**73.8**
	$\mathcal{P}_\mathcal{R}^\emptyset$	62.0	66.2	53.1
	$\mathcal{P}_\mathcal{R}^{'1}$	**76.4**	80.0	72.3
	$\mathcal{P}_\mathcal{R}^{'2}$	76.0	80.0	72.9
	$\mathcal{P}_\mathcal{R}^{'3}$	77.4	**81.1**	73.1

IV.5.2 Performance on Domain-Specific Datasets

BERT and, more broadly, EncLMs are usually pre-trained on general domain data (Wikipedia and BooksCorpus [272] for BERT), and we can ask ourselves if that impacts performances on out-of-domain datasets such as FewRel NYT and FewRel PubMed. We can see in table IV.1 that PromptORE does not see its results plummet. On the contrary, it still outperforms previous SOTA models by a large margin. SelfORE, which also relies on BERT embeddings, does not see its performance deteriorate, indicating BERT's ability to provide precise embeddings even on unseen domains[7].

The results are generally higher than with FewRel, which is explained by the fact that the two datasets contain fewer relation types and instances.

Finally, EType+ predicts a single class on FewRel PubMed because V-measure and ARI touch zero. As we have stated earlier, it is explained by the entity type schema, which is very limited as there are no *person*, *organization*, or *location* entities in this dataset.

IV.5.3 Does PromptORE "Extract" Relations?

The core of PromptORE is the prompt $\mathcal{P}_\mathcal{R}$ used by the relation encoder to embed each instance. However, one can ask if BERT really uses the text of the current instance to predict the missing token (and thus extracts information from the sentence) or if it is only using its internal knowledge, ignoring the current instance context. To answer this question, we propose to create an empty prompt $\mathcal{P}^{\emptyset}_\mathcal{R}$ where we do not input the current instance text. Its template is defined as:

$$\mathcal{P}^{\emptyset}_\mathcal{R}(\boldsymbol{e}_{head}, \boldsymbol{e}_{tail}, \boldsymbol{d}) = \text{“}\{\boldsymbol{e}_{head}\}\ \texttt{[MASK]}\ \{\boldsymbol{e}_{tail}\}.\text{”}. \tag{IV.4}$$

It is equivalent to $\mathcal{P}_\mathcal{R}$ defined in equation (IV.2), except that we have removed $\boldsymbol{d}$.

The results are shown in table IV.2. We can see that the performance for all three metrics and three datasets are much lower with $\mathcal{P}^{\emptyset}_\mathcal{R}$ compared to $\mathcal{P}_\mathcal{R}$, with an average gap of 15 % in B^3, 14 % in V-measure and 15 % in ARI. BERT benefits from the instance context to more precisely identify the relation type between the two entities.

Additionally, it is interesting to notice that even without the instance text, PromptORE still surpasses SelfORE. This clearly indicates that the method proposed by SelfORE fails to take full advantage of BERT embeddings. On the contrary, the simplicity of our approach allows PromptORE to elicit more relational information from the internal representations of BERT.

IV.5.4 Alternative Prompts

As discussed in section IV.3.1, $\mathcal{P}_\mathcal{R}$ is not necessarily the best prompt, as more relations can be expressed with a noun than a verb. We computed PromptORE performances with three alternative prompts: $\mathcal{P}'^1_\mathcal{R}$, which encourages BERT to predict a noun, $\mathcal{P}'^2_\mathcal{R}$ with the prefix proposed by Lv et al. [229], and $\mathcal{P}'^3_\mathcal{R}$ containing a prefix variant of $\mathcal{P}'^2_\mathcal{R}$. Results are shown in table IV.2.

First, we notice that $\mathcal{P}'^1_\mathcal{R}$ provides better results than $\mathcal{P}_\mathcal{R}$ in ARI, but similar performances in V-measure and B^3 for FewRel and FewRel NYT. No improvement is observed with FewRel PubMed. This result is interesting because we showed that fewer relations can be expressed with a verb than a noun, so we expected a gap in favor of $\mathcal{P}'^1_\mathcal{R}$. For example, FewRel relations *instance of*, *competition class*, *constellation*, or *operating system* cannot be expressed with a verb but are nonetheless correctly identified with $\mathcal{P}_\mathcal{R}$. BERT is weakly impacted by the apparent impossibility of predicting a meaningful and grammatically correct word in place of the `[MASK]` token.

In table IV.2, $\mathcal{P}'^2_\mathcal{R}$ and $\mathcal{P}'^3_\mathcal{R}$ achieve higher performances than $\mathcal{P}'^1_\mathcal{R}$ in the majority of the cases, while their only difference with $\mathcal{P}'^1_\mathcal{R}$ is the prefix ("In this sentence" or "We deduce that"). We can also see the impact of the prompt's wording: at first glance, both prefixes convey the same idea, but their performances differ. In fact if we replace "deduce" by "conclude" in $\mathcal{P}'^3_\mathcal{R}$ we obtain lower performances (not shown in table IV.2).

[7] As an aside, it has been observed than domain-specific fine-tuned EncLM models perform better on their respective domains (e.g., drBERT [273] for the biomedical domain). We did not employ them as the other baselines used domain-generic EncLM.

Table IV.3: Comparison of RE performances (in %) of PromptORE using different methods to estimate k. "Ideal" represents results when k is provided (same results as table IV.1). The best B^3 F1, V-measure F1, and ARI for each dataset is in **bold**.

Dataset	Method	$\hat{k}$	B^3 F1	V-measure F1	ARI
FewRel	Ideal	80	48.8	**71.8**	**43.4**
	OPTICS	571	10.8	8.5	0
	Elbow	65	**49.5**	71.2	42.2
FewRel NYT	Ideal	25	**65.1**	**78.0**	**56.9**
	OPTICS	35	33.2	29.3	1.7
	Elbow	26	64.1	77.4	56.2
FewRel PubMed	Ideal	10	**77.4**	**81.1**	**73.8**
	OPTICS	12	26.8	11.2	0.3
	Elbow	10	**77.4**	**81.1**	**73.8**

Finally, there is no consensus on the best prompt from the four proposed ones: $\mathcal{P}'^3_{\mathcal{R}}$ is the best for FewRel, $\mathcal{P}'^2_{\mathcal{R}}$ for FewRel NYT and $\mathcal{P}_{\mathcal{R}}$ for FewRel PubMed. This highlights the importance of selecting and fine-tuning prompts to maximize BERT's performances, which is indeed a significant research area for prompt-based methods [229, 241, 274, 275]. Under our fully unsupervised setting's goal, it is unfeasible to fine-tune the prompt due to the lack of labeled data. Therefore, we keep our original $\mathcal{P}_{\mathcal{R}}$ to ensure fair results.

IV.5.5 Clustering Without Knowing k

Up to now, PromptORE knows the number of relation types, k. However, as said in section IV.3.2, the most general setting is when we do not know k. We identified two methods to cluster our data without k:

1. OPTICS, a clustering algorithm based on density;
2. the elbow rule (to calculate $\hat{k}$ an estimation of k) with k-means clustering.

The results are shown in table IV.3. For the elbow rule, we first calculate multiple clusterings by varying the number of clusters. For each of these clusters, we compute the silhouette coefficient. We obtain the blue scatter plot of the figure IV.2 for FewRel. As this plot is rough, we approximate it thanks to a ridge regression with a Gaussian kernel (orange curve in figure IV.2). In a normal case, we should visually find the location of the elbow. However, the curve of figure IV.2 has a sharp elbow that is close to the curve's maximum. Thus, we propose to select $\hat{k}$ to correspond to the maximum of the silhouette score. As a side note, we noticed the same curve shape with a maximum for FewRel NYT and FewRel PubMed, so we applied the same principle.

We detect the elbow at $\hat{k} = 65$ clusters for FewRel. We obtain $\hat{k} = 26$ for FewRel NYT and $\hat{k} = 10$ for FewRel PubMed, that is, values of $\hat{k}$ nearly identical to the real number of relations.

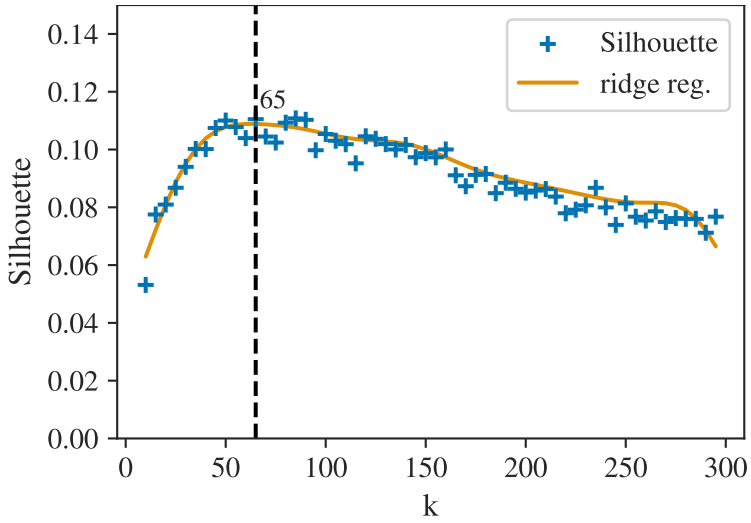


Figure IV.2: Silhouette curve computed to estimate the number of clusters $\hat{k}$ with the elbow rule. Due to the roughness of the curve, it was approximated using a ridge regression with a Gaussian kernel. The estimation of the number of relation types is at the curve's maximum ($\hat{k} = 65$).

Quantitative Results OPTICS sets $\hat{k}$ at 571 (see table IV.3), far from the optimal $k = 80$ for FewRel. It translates into poor performances compared to PromptORE when we know k. We also note that OPTICS is very slow during training (approximately 6 h compared to 5 min with k-means). On the other side, results are much more satisfactory with the elbow rule, with a slight decrease in ARI but equivalent performances in B^3 and V-measure. Training time is also much more reasonable with approximately 1 h. We make the same conclusion when we look at FewRel NYT and FewRel PubMed.

We can conclude that, at least on our three datasets, the Elbow Rule efficiently finds a correct estimation of k. Finally, it is interesting to see that PromptORE with the elbow rule widely surpasses previous state-of-the-art approaches (see table IV.1), with a gap of 15 % – 25 % in B^3 and V-measure and 17 % – 30 % in ARI. We demonstrate that it is possible to remove the dependency on all hyperparameters, including k, and still achieve state-of-the-art results.

Qualitative Analysis of the Clustering From table IV.3, the elbow rule finds $\hat{k} = 65$ instead of 80 for FewRel. This means the clustering is not ideal: some clusters must contain multiple relation types. The confusion matrix between the true relation types and the clusters predicted by PromptORE is shown in figure IV.3. It is evidently not square as the number of clusters $\hat{k}$ does not equal the number of relations k. We reorganized the axes to find a logical representation of the confusion matrix, as initially, there was no link between the cluster IDs and the relation types. To do that, we employ a method similar to the algorithm described in appendix A.

On this confusion matrix, we notice that some clusters are not pure: they contain multiple relations (e.g., clusters 11, 19, 29-32, 50, or 55). The main observation is, nevertheless, that the matrix possesses a clear diagonal, meaning that PromptORE effectively distinguishes the vast majority of the relation types while training in a fully unsupervised setting.

We also see that clusters seem relatively complete: seldom do clusters share the same relations (except for clusters 32 and 42 or 40, 50, and 55).

Table IV.4: Relation types composing four randomly sampled impure clusters predicted by PromptORE tested on FewRel with the elbow rule. Relation types are displayed in descending proportion order.

Cluster	Relation Types	
11	47 %	language of film or TV show
	40 %	language of work or name
	7 %	country of origin
	6 %	*other*
19	48 %	mountain range
	27 %	located in physical feature
	15 %	located in or next to body of water
	4 %	located in the administrative territorial entity
	6 %	*other*
32	27 %	screenwriter
	27 %	director
	20 %	after a work by
	13 %	characters
	9 %	composer
	4 %	*other*
50	25 %	sibling
	21 %	father
	19 %	child
	18 %	mother
	17 %	spouse

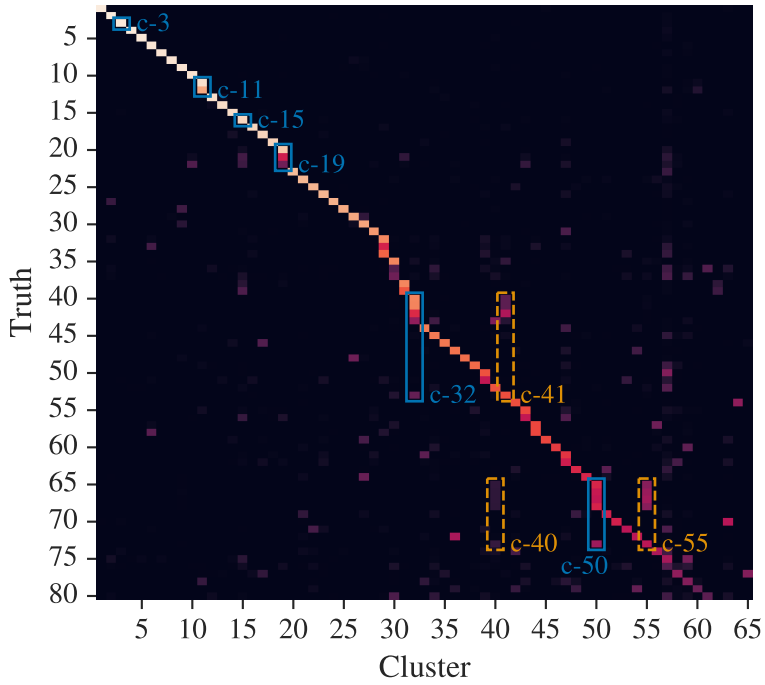


Figure IV.3: Confusion matrix of PromptORE tested on FewRel when $\hat{k}$ is estimated with the elbow rule. Columns and rows were reordered using a method similar to the algorithm described in appendix A (as there is no direct link between the cluster IDs and the relation types). The main relation types of some clusters are highlighted.

As some clusters contain multiple relation types, we find it interesting to check whether the types that compose each cluster are semantically linked. We randomly sample four clusters that contain multiple relation types. Results are shown in table IV.4. For these four clusters, we can see that the types are indeed semantically close within a cluster:

- for cluster 11, they are linked to the language of an artistic work (*country of origin* is related to the language);
- for cluster 19, to geographical location;
- for cluster 32, to artistic creation;
- for cluster 50, to family relationship.

Even though these clusters are not optimal from the FewRel annotation point of view, they are semantically coherent. We could even argue that cluster 11 makes more sense than the initial labeling, which divided this cluster into relation types with generic types (*language of work or name*) and a relation type specific to cinema and movies (*language of film or TV show*).

In conclusion, this qualitative analysis shows that combining PromptORE with the elbow rule efficiently discovers semantically consistent clusters, which are very close to true relations. This result in itself is very impressive: without any prior knowledge about the relation types (name, description, or even number), PromptORE identifies most of them precisely and, in any case, finds a coherent typing scheme. Finally, the scheme proposed by PromptORE and its

Table IV.5: Most frequent predicted tokens for three clusters identified by PromptORE with the elbow rule for FewRel. The relation types composing these clusters are also displayed in descending proportion order.

Cluster	Predicted token	Relation types	
3	,	99 %	platform
	for	1 %	*other*
	supports		
15	:	79 %	contains administrative territory
	borders	5 %	located in administrative entity
	,	3 %	located in physical feature
	surrounds	7 %	*other*
50	,	25 %	sibling
	.	21 %	father
	married	19 %	child
		18 %	mother
		17 %	spouse

clear diagonal is all the more stunning given the high number of classes (80), which renders this classification non-trivial. In our opinion, it demonstrates the quality and value of using EncLM prompting with clustering to classify relation types in an unsupervised setting. We believe this technique can also be applied to other information extraction tasks (see chapter V).

IV.5.6 Analysis of $\mathcal{P}_\mathcal{R}$ Prompt Predictions

In section IV.5.4, we found a very little performance difference between $\mathcal{P}_\mathcal{R}$ and $\mathcal{P}'^1_\mathcal{R}$, while the majority of relations cannot be written using a verb. To go further, it would be interesting to check which tokens/verbs best describe the clusters identified by PromptORE with $\mathcal{P}_\mathcal{R}$. To do that, we use the relation embeddings (computed by our relation encoder) and predict the masked tokens (represented by `[MASK]`) with the MLM head of BERT. By iterating for every instance located in one cluster, we can find the most frequent tokens that describe it. We apply this method on the clusters identified by PromptORE and elbow rule (see figure IV.3). The results on three selected clusters are displayed in table IV.5.

In most cases, the predicted names are not clear enough to qualify the relation type corresponding to the cluster. Nevertheless, the names give clues to identify the general theme of the relation type ("married" indicates a family-centered relation type, "borders", and "surrounds" a geographic topic).

The table IV.5 also shows the major limitation of $\mathcal{P}_\mathcal{R}$: when we look at cluster 50, *spouse* is represented by "married", but *sibling*, *father*, *child*, and *mother* relation types are not brought to light with the predicted names. Indeed, these four types cannot be written using a single verb; by not finding satisfactory names, BERT defaulted to predicting punctuation tokens. We

reach the same conclusion for cluster 15, where BERT predicts punctuation tokens.

Finally, this observation gives us interesting insights about the behavior of our relation encoder. Intuitively, we could think that PromptORE would have poor results with relation types that cannot be written using a verb. On the contrary, we found that results were close between $\mathcal{P}_{\mathcal{R}}$ and $\mathcal{P}'^{1}_{\mathcal{R}}$ (see table IV.2). At the same time, we notice that cluster 3 (table IV.5) is very pure, yet its most predicted name is ",", a token that is furthermore shared among the two other clusters of table IV.5. This indicates that BERT can encode a very expressive embedding of the current relation instance that allows precise clustering but cannot be translated into real words. This is supported by the fact that PromptORE identifies three different clusters with punctuation as their most frequent tokens (table IV.5). It is reassuring to see that complex prompts are not required to represent many relation types effectively. It does not undermine the importance of prompt-tuning (when possible): they impact model performances as shown in table IV.2.

IV.6 Conclusion

In this chapter, we introduced PromptORE, our unsupervised RE model. The primary motivation of PromptORE is the weakness of previous state-of-the-art models, which heavily rely on hyperparameters, in particular, to prevent overfitting. Their respective authors do not explain how to adjust them without using an annotated validation dataset, which is incompatible with the unsupervised scenario. To solve this shortcoming, our proposed approach leverages and adapts the prompt-tuning paradigm to work under an unsupervised setting. This allows us to create a flexible relation typing model that does not require training or fine-tuning, thus removing nearly all hyperparameters. We also employ a simple heuristic to estimate the number of relation types.

Experiments on one general and two domain-specific datasets show that PromptORE widely surpasses previous state-of-the-art methods (with a gap of 18 % – 19 % in B^3, V-measure, and ARI) while being simpler and not needing any hyperparameter tuning. The qualitative analysis of the confusion matrices demonstrates that PromptORE identifies most relation types without prior knowledge and provides a semantically coherent typing scheme. The results are auspicious, given the relatively high number of classes in FewRel (80), showing that PromptORE can identify precisely various relation types.

Finally, PromptORE is an experimental validation of the feasibility of realistic low-resource (unsupervised) and open-world relation extraction. The next logical step is to adapt and enhance the PromptORE method to the second primary task of information extraction: named entity recognition.

V CITRUN

Cross-Domain Transfer-Learning for Unsupervised Named Entity Recognition

Unsupervised and zero-shot Named Entity Recognition (NER) aims to extract and classify entities in documents from a target domain $\mathcal{D}_T$ without annotated data, a very low-resource setting. While zero-shot NER approaches yield impressive outcomes, they operate under the assumption that all entity types found in $\mathcal{D}_T$ are predefined and known (closed-world hypothesis). Unsupervised NER tackles this limitation, but existing models suffer from the same reliance on hyperparameters as unsupervised RE baselines (chapter IV), making them unusable in a real-world scenario.

To address these shortcomings, we introduce CITRUN. This unsupervised NER model does not need annotations in $\mathcal{D}_T$ (in particular to adjust hyperparameters) and does not require knowledge of the target entity types or their number. To do that, we extend the typing approach developed with PromptORE (described in chapter IV) for the NER task and use it to structure the extracted entities in entity types. We also propose to use contrastive learning to refine entity embeddings and elicit entity types more precisely. Results on 13 domain-specific datasets show that CITRUN significantly outperforms unsupervised LLM prompting and is competitive with the latest zero-shot models. Qualitative analysis shows that CITRUN effectively groups entities into semantically meaningful clusters resembling the true entity types.

The source code of CITRUN and the unsupervised baselines is available in a public repository[1] and distributed under an open-source license.

Contents

[1] Available at https://github.com/alteca/CITRUN.

V.1 Introduction

Named Entity Recognition (NER) is a fundamental NLP task that aims to identify entities in text and classify them into entity types. It is the logical next step after the success of PromptORE with unsupervised RE. The main objective of this chapter is to transfer and improve the results of PromptORE on the NER task.

Similarly to RE, NER has primarily been approached as a supervised task [39, 113], which presents challenges in specific domains (e.g., scientific, biomedical) where large labeled corpora may not be readily available. As a consequence, interest in low-resource and few-shot NER has risen [116, 276, 277], especially since the emergence of Encoder-Only Language Models (EncLM) such as BERT [6]. However, these approaches still require annotation.

Zero-shot NER aims to alleviate this constraint. Recent models typically transfer knowledge from a source domain $\mathcal{D}_S$ to a target domain $\mathcal{D}_T$ where no annotated data is available [31, 37]. Although they do not require labels in $\mathcal{D}_T$, they assume a closed-world hypothesis, where entity types are known in advance, making them inapplicable in exploratory scenarios or novelty detection use cases. To solve this shortcoming, tentatives for unsupervised NER have been proposed. However, they suffer from the same weakness highlighted with PromptORE: a heavy reliance on hyperparameters that cannot be adjusted without labeled data.

Therefore, we introduce **CITRUN**, our "**C**ross-Doma**I**n **TR**ansfer-Learning for **U**nsupervised **N**amed Entity Recognition" model. CITRUN follows a strict unsupervised setting: no labeled data in $\mathcal{D}_T$, no knowledge of the entity types $\mathcal{E}_T$ (including their number), and no hyperparameter tuning dependant on labeled data from $\mathcal{D}_T$.

To do that, we adapt the methodology developed with PromptORE to unsupervised NER for the typing part and extend it in several aspects. We propose to use annotated data from a source

domain $\mathcal{D}_S$ to train CITRUN and transfer it to the target domain $\mathcal{D}_T$[2], envisioning that CITRUN will learn generalizable patterns from $\mathcal{D}_S$ that apply to $\mathcal{D}_T$. We include mention detection in the spectrum of CITRUN, whereas PromptORE previously ignored the relation detection part. For entity typing, we propose a novel embedding refinement approach based on contrastive learning to isolate entity types more effectively in $\mathcal{D}_T$. Additionally, we rework the cluster estimation algorithm to make its computation more effective. To summarize our main contributions:

- We propose CITRUN, an unsupervised NER model that extracts and classifies entities from a target domain $\mathcal{D}_T$ without annotations in $\mathcal{D}_T$, without knowing the target entity types $\mathcal{E}_T$, nor their number $|\mathcal{E}_T|$ (section V.3). We generalize PromptORE's prompting and clustering approach to classify entities into entity types (section V.3.2).
- We extend PromptORE to use contrastive learning to elicit entity types more precisely and propose a logarithmic complexity class algorithm to estimate the number of clusters (section V.3.2.3).
- Experimental results on 13 domain-specific datasets show that CITRUN surpasses LLM-based unsupervised NER and performs comparably to the more supervised zero-shot and few-shot state-of-the-art models (section V.5). Qualitative analysis highlights that CITRUN organizes entities in semantically coherent clusters close to true entity types (section V.5.8).

V.2 Related Work

V.2.1 Few-Shot & Zero-Shot Named Entity Recognition

Most of the few-shot and zero-shot models assume the availability of labeled data in a source domain $\mathcal{D}_S$ and try to learn from $\mathcal{D}_S$ and transfer to $\mathcal{D}_T$. $\mathcal{D}_S$ can be a manually annotated dataset [23, 277], a distantly labeled dataset [278], or a synthetically generated dataset [31, 66, 124]. Recent approaches are divided into two families: 1. two-stage NER, and 2. one-stage or integrated NER.

Two-stage approaches split NER into Mention Detection (MD) and Entity Typing (ET) [277, 279, 280]. MD aims to identify spans of $\boldsymbol{d}$ that are entity mentions, and ET classifies the type of each extracted mention. Integrated models combine MD and ET in one step, the motivation being to reduce cascading errors [23, 31, 278, 281, 282]. In practice, both paradigms attain state-of-the-art results [277, 281]. Until recently, most approaches relied on EncLMs. We see now the rising use of Large Language Models (LLM) in these two low-resource settings [23, 31], where LLMs particularly shine [155].

Mention Detection (MD) Few-shot and zero-shot approaches follow architectures similar to supervised models for MD. They usually implement span-based extractors [66, 116, 283], although BIO sequence labeling is still used [277, 280]. These extractors are trained in a

[2] $\mathcal{D}_S$ that is different from $\mathcal{D}_T$. For instance, the source domain can be a news dataset (such as CoNLL-2003 [39]), and the target domain a biomedical dataset (e.g., i2b2 [209]).

supervised fashion on $\mathcal{D}_S$ entities. The challenge involves transferring the learned patterns from $\mathcal{D}_S$ to $\mathcal{D}_T$ mentions. BIO sequence labeling classifies each token t in $\boldsymbol{d}$ as either B (first token of a mention), I (second or following token of a mention), or O (not a mention). A decoding algorithm then reconstructs the boundaries of the mention using the predicted classes. Greedy algorithms are used, especially with recent language models [277]. Conditional random fields [284] are extensively employed to improve decoding. The main weakness of BIO is that it cannot predict nested entities. This is the main motivation for span-based extractors.

In general, span-based extractors score each possible span in $\boldsymbol{d}$ and determine the true entity mentions [113, 116]. To do that, they compute *start of span* and *end of span* vector representations (usually embeddings of the first and last tokens of the candidate span). Zhong et al. [113] concatenate the *start* and *end* embeddings and use them in a perceptron that scores the candidate span. Wang et al. [116] use bilinear layers to replace the perceptron, allowing more efficient computations compared to Zhong et al. [113]. Span-based approaches suffer from the quadratic number of possible spans, making scoring the candidate spans expensive for long documents. Dobrovolskii [58] tries to overcome this problem with a hybrid approach. First, each word in $\boldsymbol{d}$ is classified as an entity head or not. An entity head is the main word of an entity; Dobrovolskii considers the head to be the root of the syntactic subtree of the mention. This ingenuity allows him to lower the quadratic complexity to a linear complexity. Once the entity heads are identified, the boundaries of each mention are determined using a convolutional neural network. Finally, Zaratiana et al. [285] propose to adapt conditional random fields for span-based extractors to enforce non-overlapping spans.

Entity Typing (ET) The general principle is to compute a vector representation of the extracted entity mentions (entity embeddings) and compare them to those of the exemplars (few-shot) or the target entity types (zero-shot and few-shot). Zhang et al. [279] propose to use the k-nearest neighbors with the few-shot exemplars to identify the type. Prototypical networks [237] are generally preferred to classify entities [116, 277, 283]. The entity-type prototypes are computed using the exemplars.

Entity embeddings are computed by aggregating the EncLM embeddings of the individual tokens composing the mention in the case of a BIO extractor [277] or by using the span representation used by the span extractor [116]. Shen et al. [278] and Ding et al. [286] explore prompting techniques with EncLM (using the `[MASK]` token) as an alternative to generate entity embeddings.

Meta-learning [287] is employed to enhance the efficacy of transfer learning [280]. The idea is to generate large amounts of few-shot episodes using the annotated data of $\mathcal{D}_S$ (the set of entity types of $\mathcal{D}_S$); each episode contains a subset of $\mathcal{E}_S$, randomly selected few-shot exemplars associated to $\mathcal{E}_S$, and test documents to compute the performance. Then, the model is trained on the episodes to achieve the best transfer in the smallest fine-tuning steps possible (hence the meta-learning term). This allows fine-tuning even on the limited few-shot exemplars, as the model is adapted to converge quickly and reliably.

Finally, Liu et al. [276] and Mahapatra et al. [288] explore the effectiveness of adapting EncLM embeddings to the target domain. They employ large amounts of unannotated documents of $\mathcal{D}_T$ and fine-tune BERT weights using a masked language modeling task. Empirically,

they observe a link between a decrease in perplexity and an increase in NER performances. Mahapatra et al. [288] decrease the training time required for domain adaptation by filtering the unannotated documents of $\mathcal{D}_T$ to keep those more aligned to the actual documents where entities are to be extracted.

Large Language Models Very recently, Large Language Models (LLMs) [16, 289] have been successfully applied to few-shot and zero-shot NER and have state-of-the-art results on the zero-shot setting.

First, "raw" prompting obtains impressive results compared to previous works [146, 151, 257]. Wang et al. [146] and Ye et al. [257] requires few-shot exemplars to specify the output format. Wei et al. [151] (ChatIE) propose a multi-turn framework that works in a zero-shot setting (without the need for exemplars). Surprisingly, they reverse the usual MD and ET steps order. Indeed, they first ask the LLM which entity types are present in the document (given a predefined list of entity types). In subsequent turns, they ask the LLM about the entities associated with each entity type. The weakness of their work is the multiple turns required to analyze a document, which is expensive when using the APIs of the largest LLMs.

Sainz et al. [23], Zhou et al. [31], and Wang et al. [143] explore the idea of fine-tuning small LLMs [16, 290, 291] on manually or synthetically labeled datasets. In doing so, they create NER-specialized LLMs with better performances than generalist LLMs while being much smaller. Zhou et al. [31] propose to annotate documents from the Pile corpus [159] using GPT-3.5 (they call this dataset Pile-NER) and fine-tune Vicuna [290] on it. Their UniNER model achieves better performances than GPT-3.5 in a zero-shot context. Additionally, fine-tuning using a large amount of synthetic data allows them to specify a custom JSON format that UniNER follows reliably. GoLLIE [23] uses Code-Llama [292] as its backbone and is fine-tuned on manually labeled datasets from the news and biomedical domains. Sainz et al. [23] follow a "Python class" scheme, where each entity type is specified as a Python class with a name, a description, and a few examples. They find empirically that description and exemplars metadata positively impact GoLLIE performances.

Regarding the prediction format, most of the approaches follow a surface form extraction scheme [23, 31, 151, 257], except GPT-NER [146]. The models output only the mention text, and a subsequent algorithm is required to localize the mention in the document. The output format is generally JSON, but Sainz et al. [23] use Python code, allowing them to add metadata elegantly in comments (description and exemplars). GPT-NER [146] proposes a sequence labeling scheme. It asks the LLM to repeat the input document, with a special markup delimiting the boundaries of mentions: `@@` as the opening tag and `##` as the closing tag. This format removes the dependency on a decoding algorithm, as the detected mentions are localized in the document by design. However, it is incompatible with a zero-shot setting, as in-context exemplars are required to describe the output format.

Finally, Zaratiana et al. [66] (GliNER) fine-tune EncLM embeddings (DeBERTa v3 [293]) on the GPT-3.5 generated annotations of Pile-NER [31]. They implement a span-based extractor for MD coupled with a method similar to prototypical networks for ET. They obtain very competitive results compared to the much larger fine-tuned LLMs UniNER [31] and GoLLIE

[23]. Today, this model represents the best balance between the flexibility of LLM-based zero-shot NER and the relatively small number of parameters of EncLM embeddings.

V.2.2 Unsupervised Named Entity Recognition

Historically, unsupervised NER implemented rules and patterns-based models [294, 295]. However, they were specific to a small set of entity types, hindering the discovery of unspecified types. In fact, the most recent unsupervised NERs suffer from the same problem and require prior knowledge of the target entity types [35, 37, 38, 296]. Formally, they are zero-shot models and not completely unsupervised approaches.

Jia et al. [35], Peng et al. [37], and Liu et al. [296] propose to generalize the transfer-learning from $\mathcal{D}_S$ to $\mathcal{D}_T$ setting used in the zero-shot setting. They train entity-type-specific models based on BERT embeddings, which are merged together in a mixture of experts. They must know the target entity types beforehand and need access to labels for each entity type (from a different domain, but still annotations). CycleNER [38] proposes a seq-to-seq model with a double translation mechanism between text and entities. It comprises two models: S2E translating the document into its list of entities and E2S generating the text from a list of entities. The two models are trained jointly, and S2E is kept for predictions. CycleNER also must know the target entity types in advance and requires lists of entities from $\mathcal{D}_T$.

Only UNER [297] is compatible with an unsupervised scenario. UNER uses clustering for MD and employs self-learning with auto-encoders for ET. UNER is subject to drifting (as it relies on self-learning) and requires careful hyperparameter tuning (number of training steps, learning rate, etc.) to prevent catastrophic performance drops. Unfortunately, UNER lacks a detailed explanation of how these hyperparameters are adjusted unsupervised. As an aside, it is interesting to notice that unsupervised NER models suffer from the same hyperparameter tuning critique we emitted for unsupervised RE and the same use of unstable self-learning paradigms that are particularly difficult to optimize.

V.3 Description of CITRUN

CITRUN aims to extract and type entity mentions from documents $\boldsymbol{d}$ of $\mathcal{D}_T$ in an unsupervised setting. Given $\boldsymbol{d}$, the objective is to identify the spans $\boldsymbol{m} = [t_i, ..., t_j] \in \boldsymbol{d}$ that are entity mentions, and classify the type $\mathfrak{e} \in \mathcal{E}_T$ for each $\boldsymbol{m}$. CITRUN assumes no prior knowledge of $\mathcal{D}_T$. It does not have access to:

- annotated documents of $\mathcal{D}_T$,
- the set of entity types $\mathcal{E}_T$,
- the number of entity types $|\mathcal{E}_T|$.

Similarly to recent zero-shot and few-shot models [31, 66, 278], CITRUN is built upon a cross-domain transfer-learning scheme. The general idea is to learn the NER task on a source domain $\mathcal{D}_S$, where annotated data is available, and transfer it to $\mathcal{D}_T$. $\mathcal{D}_S$ differs from $\mathcal{D}_T$

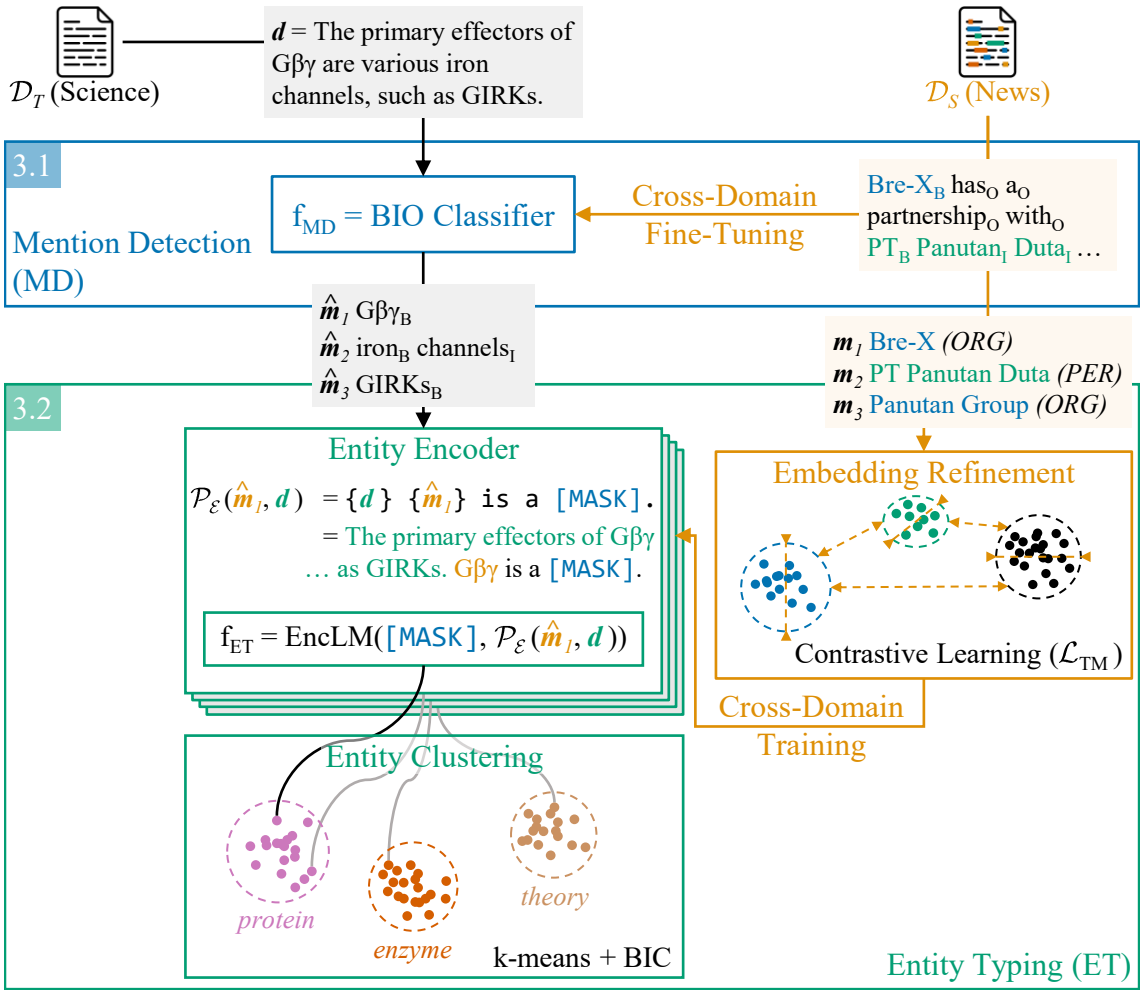


Figure V.1: Overall architecture of CITRUN.

stylistically, semantically, and/or from the entity type perspective ($\mathcal{E}_S \neq \mathcal{E}_T$). We go beyond zero-shot and few-shot approaches by not predefining $\mathcal{E}_T$.

As shown in figure V.1, CITRUN follows a two-step process, with:

1. Mention Detection (MD). It identifies the spans $\boldsymbol{m}$ of $\boldsymbol{d}$ that are entity mentions.
2. Entity Typing (ET). It classifies the type $\mathfrak{e}$ for each extracted mention. In practice, CITRUN finds clusters of entities with the same type $\mathfrak{e}$.

V.3.1 Mention Detection (MD)

MD identifies entity mentions $\boldsymbol{m}$ for a given document $\boldsymbol{d}$.

As we have presented in the previous section, two main prediction paradigms exist for MD: BIO sequence labeling extractors [66, 116, 283], and span-based extractors [277, 280]. In general, span-based extractors achieve slightly better results than BIO models in supervised and low-resource settings [66, 113]. We choose to formulate MD as a BIO sequence labeling, classifying each $t_i \in \boldsymbol{d}$ as B (first token of an entity), I (second or following token of an entity), or O (not an entity). We employ a BIO extractor due to its lower expressivity and complexity than span-based models, expecting it to lead to better generalizability on unseen domains and new entity types.

We employ EncLM embeddings, coming from pre-trained language models such as BERT [6], combined with a linear classifier:

$$\mathrm{f}_{\mathrm{MD}}(t_i, \boldsymbol{d}) = \sigma(\mathrm{EncLM}(t_i, \boldsymbol{d})\boldsymbol{W} + \boldsymbol{b}), \tag{V.1}$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ are learned weights, $\mathrm{EncLM}(t_i, \boldsymbol{d})$ is the EncLM embedding of t_i in the context of $\boldsymbol{d}$, and σ is the softmax function. f_{MD} is fine-tuned (EncLM weights, $\boldsymbol{W}$ and $\boldsymbol{b}$) on annotated documents from $\mathcal{D}_S$.

In fact, MD is the primary motivation for annotated data. On the one hand, we demonstrated in the chapter IV the possibility of precisely typing relations without needing training data. Thus, using the same principle for ET seems reasonable, eliminating the need for labeled documents for this step. On the other hand, the only MD model that works unsupervised relies on self-learning [297]. Yet, self-learning is known to be subject to drifting when overtrained. Preventing drifting requires careful hyperparameter tuning (especially the number of training steps and the learning rate). Luo et al. [297] do not explain how to adjust them without external annotated $\boldsymbol{d}$ from $\mathcal{D}_T$. As a result, we propose to use annotated documents from $\mathcal{D}_S$ to train MD in a supervised fashion (but cross-domain) to diminish the risk of unstable results. As a side note, annotations for $\mathcal{D}_S$ may come from manually labeled datasets, distantly annotated datasets [278], or synthetically generated data [31]. In this chapter, we train CITRUN on manually labeled and synthetically generated datasets (see section V.5.2).

V.3.2 Entity Typing (ET)

ET classifies the entities previously extracted with MD. In an unsupervised setting, the objective is to group entities with the same entity type $\mathfrak{e} \in \mathcal{E}_T$. As shown in figure V.1, ET comprises three

modules. ET is the application of the principles of PromptORE for entity typing and contains several improvements regarding the estimation of the number of clusters (section V.3.2.2) and the use of $\mathcal{D}_S$ to refine embeddings (section V.3.2.3).

V.3.2.1 Entity Encoder

The first module of ET is the entity encoder, which computes a vector representation (or entity embedding) of the current entity. We want this embedding to represent the entity type: two entity mentions $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ with close embeddings should have the same type $\mathfrak{e}$. Conversely, two mentions with different $\mathfrak{e}_1$ and $\mathfrak{e}_2$ should have remote entity embeddings. To encode entities, we use the same prompting with EncLM technique as PromptORE [33] (section IV.3.1). We also choose the simplest prompt template possible:

$$\mathcal{P}_{\mathcal{E}}(\boldsymbol{m}, \boldsymbol{d}) = \text{“}\{\boldsymbol{d}\}\,\{\boldsymbol{m}\}\texttt{ is a [MASK].}\text{”}. \tag{V.2}$$

For instance:

$\boldsymbol{d}$ = “The primary effectors of Gβγ are various iron channels, such as GIRKs.”,
$\boldsymbol{m}$ = “Gβγ”,
$\mathcal{P}_{\mathcal{E}}(\boldsymbol{m}, \boldsymbol{d})$ = “The primary effectors of Gβγ are various iron channels, such as GIRKs. Gβγ is a [MASK].”.

The entity representation is then computed as the embedding of [MASK] in the context of the prompt $\mathcal{P}_{\mathcal{E}}(\boldsymbol{m}, \boldsymbol{d})$:

$$\mathrm{f}_{\mathrm{ET}}(\boldsymbol{m}, \boldsymbol{d}) = \mathrm{EncLM}(\texttt{[MASK]}, \mathcal{P}_{\mathcal{E}}(\boldsymbol{m}, \boldsymbol{d})). \tag{V.3}$$

V.3.2.2 Entity Clustering

The second module is the entity clustering. We follow a principle similar to the relation clustering module of PromptORE (section IV.3.2). Once all entities extracted in $\mathcal{D}_T$ are encoded using the previous module, we cluster the embeddings to identify groups of entities that are close, relative to the Euclidian distance, and thus expected to have the same type $\mathfrak{e} \in \mathcal{E}_T$. We use the simple k-means algorithm [258, 259]. Since the number of entity types is unknown, the number of clusters must be estimated.

Improving PromptORE's Elbow Rule With PromptORE, we employed the elbow rule method with the silhouette coefficient. Although this method provides good estimations, it has the drawback of being a visual approach. Automatic approaches have been proposed to determine the elbow but are less precise than the human eye [268, 269]. In the tested datasets, we could approximate the silhouette coefficient with a relatively smooth curve that had a clear maximum close to the elbow, making the detection easy (by finding the extremum), but this is not always the case. In particular, this is not true with CITRUN, which has a rough silhouette curve, making the previous heuristic inapplicable.

To tackle this limitation, it is interesting to notice that k-means can be seen as a simplification and approximation of a spherical Gaussian Mixture Model (GMM) [298]. The main difference

resides in cluster membership: with k-means, each point belongs only to one cluster (Dirac probability distribution), whereas GMM produces soft-clustering assignments. One approach to estimate the number of clusters of a GMM is to fix an upper bound K, compute a clustering for each $k, 2 \leq k \leq K$, compute the Bayesian Information Criteria (BIC) [299] for each clustering and select $\hat{k}$ that minimizes BIC. The major advantage of this approach is that it does not require to locate the elbow but rather a minimum, which is easier to find. This is because BIC measures the quality of the clustering (similar to silhouette) and adjusts it relative to the complexity of the model. Indeed, when looking at the right-hand side of equation (V.4), the left term measures the quality of the fit, while the right part estimates the complexity of the model. We propose to apply this same procedure to estimate the number of clusters with k-means, using the k-means BIC formula of Onumanyi et al. [269]:

$$BIC = n \ln(\frac{RSS}{n}) + k \ln(n), \tag{V.4}$$

$$RSS = \sum_{0 \leq i < n} (\mathrm{f_{ET}}(\boldsymbol{m}_i, \boldsymbol{d}_i) - \boldsymbol{c}_i)^2, \tag{V.5}$$

with n the number of entity mentions $\boldsymbol{m}_i$ extracted by MD, $\boldsymbol{d}_i$ the document containing $\boldsymbol{m}_i$ and $\boldsymbol{c}_i$ the centroid of the cluster containing $\boldsymbol{m}_i$. We call this procedure *brute force cluster estimation*. This is the main approach we employ during CITRUN's evaluation.

Reducing Computational Complexity One constraint of the previous approach is that it requires to compute a clustering for each $2 \leq k \leq K$, which is computationally expensive. Empirically, we find the BIC curve for ET to be smooth, globally convex, and with a single minimum (see figure V.7). This was observed for the 13 $\mathcal{D}_T$ datasets used during evaluation (see section V.4.2), different EncLM embeddings, and every variation of CITRUN. With this experimental observation, finding the global minimum BIC without testing every possible k is possible. One such method is the ternary search. We propose implementing it and call it *ternary search cluster estimation*. The ternary search follows an iterative approach, with each cycle being:

1. In input, we have a lower bound k_{min} and an upper bound k_{max} for the number of clusters.
2. Select k_1 and k_2 such as they divide the search space between k_{min} and k_{max} in thirds.
3. For k_1 and k_2, compute the clustering and calculate the BIC.
4. If k_1 has a lower BIC than k_2, then $k_{max} = k_2$, else $k_{min} = k_1$.

The cycle is repeated until $k_{max} = k_{min}$. At each cycle, the search space is reduced by a third, giving a logarithmic complexity of $\mathcal{O}(\log_3(K) \cdot \text{k-means})$, compared to $\mathcal{O}(K \cdot \text{k-means})$ for the brute force method.

In practice, three improvements can be made. First, if the lowest BIC is at k_{min}, we set $k_{max} = k_1$; and conversely, if the lowest BIC is k_{max}, we set $k_{min} = k_2$. It allows the elimination of two-thirds of the search space in one cycle.

Secondly, we propose to remove the need to fix an upper bound K by providing a first estimate $K = \sqrt{n}$ and allowing the ternary search to increase K if the minimum BIC is located after it. During the first cycle, if the lowest BIC is located at k_{max}, instead of updating k_{min}, we set $k_{max} = k_{max} + \frac{k_{max} - k_{min}}{3}$. This move is possible for the following cycles until the lowest BIC is not at k_{max}.

Finally, the BIC curve is not completely smooth locally. To improve the minimum estimation accuracy, when k_{min} and k_{max} are close (e.g., $k_{max} - k_{min} \leq 5$), we compute every clustering for $k_{min} \leq k \leq k_{max}$ and select $\hat{k}$ with the lowest BIC.

V.3.2.3 Embedding Refinement (ER)

ET is not trained using labeled documents. However, since MD uses labeled data in $\mathcal{D}_T$, we can also employ them for ET to isolate entity types more clearly during the clustering. Contrastive learning has been applied for this purpose in the context of low-resource NER [281, 282]. The objective is to bring entity mentions of the same type closer and move away entities of different types by optimizing EncLM representations. Existing models apply contrastive learning on the annotated data of $\mathcal{D}_T$, which we do not have. As a result, we propose optimizing the contrastive loss on entity mentions of $\mathcal{D}_S$, anticipating that the reorganized embedding space will also benefit mentions in $\mathcal{D}_T$.

We implement the widely used triplet margin loss $\mathcal{L}_{\text{TM}}$ [300]. $\mathcal{L}_{\text{TM}}$ considers entity mentions triplets $(\boldsymbol{m}^a, \boldsymbol{m}^+, \boldsymbol{m}^-)$. $\boldsymbol{m}^a$ is called the anchor. The positive mention $\boldsymbol{m}^+$ has the same type as the anchor $\boldsymbol{m}^a$, and the negative mention $\boldsymbol{m}^-$ has a different type than $\boldsymbol{m}^a$. The objective of $\mathcal{L}_{\text{TM}}$ is to ensure that $\boldsymbol{m}^+$ is closer to $\boldsymbol{m}^a$ than $\boldsymbol{m}^-$ up to a certain margin. We have:

$$\mathcal{L}_{\text{TM}}(\boldsymbol{m}^a, \boldsymbol{m}^+, \boldsymbol{m}^-) = \max[0, \text{d}(\boldsymbol{m}^a, \boldsymbol{m}^+) - \text{d}(\boldsymbol{m}^a, \boldsymbol{m}^-) + 1] \tag{V.6}$$

with $\text{d}(\boldsymbol{m}^a, \boldsymbol{m}^+)$ the Euclidian distance between $\text{f}_{\text{ET}}(\boldsymbol{m}^a, \boldsymbol{d})$ and $\text{f}_{\text{ET}}(\boldsymbol{m}^+, \boldsymbol{d})$. f_{ET} weights are fine-tuned on entities of $\mathcal{D}_S$ using $\mathcal{L}_{\text{TM}}$. We fix the $\mathcal{L}_{\text{TM}}$ margin at 1. Empirically, we have not found that the margin significantly impacted the performances.

Contrary to usual EncLM fine-tuning, a larger batch size is beneficial with contrastive learning [301], as it helps regularize the embedding space reorganization. The limiting factor to increase the batch size with ET is entity encoding. For each triplet $(\boldsymbol{m}^a, \boldsymbol{m}^+, \boldsymbol{m}^-)$, three prompts $\mathcal{P}_{\mathcal{E}}$ need to be encoded. This comes with a substantial GPU footprint, hindering large batch sizes. To mitigate this issue, we change the perspective and consider batches of entity mentions instead of batches of triplets. Each mention $\boldsymbol{m}$ is associated with the document $\boldsymbol{d}_{\boldsymbol{m}} \in \mathcal{D}_S$ in which it appears and its type $\mathfrak{e}_{\boldsymbol{m}} \in \mathcal{E}_S$. We encode one prompt for each entity. Then, we find all valid triplets inside the batch, respecting the condition $(\mathfrak{e}_{\boldsymbol{m}^+} = \mathfrak{e}_{\boldsymbol{m}^a}) \wedge (\mathfrak{e}_{\boldsymbol{m}^-} \neq \mathfrak{e}_{\boldsymbol{m}^a})$. We can encode 128 entities per batch in our experimental setup. Without this optimization, one batch comprises 42 triplets, and $\mathcal{L}_{\text{TM}}$ does not converge. With this optimization, one batch contains, on average, more than 100,000 valid triplets.

System Message: You are a helpful information extraction system.

Prompt: Given a passage, your task is to extract all entities and identify their entity types. The output should be in a list of tuples of the following format: [("entity 1", "type of entity 1"), ...].

Passage: {d}

Figure V.2: Unsupervised prompt used by Zhou et al. [31] to annotate Pile-NER. It also corresponds to the prompt of UniNER Uns (GPT-3.5).

V.4 Experimental Setup

V.4.1 Baselines

Luo et al. [297] did not release the source code of UNER, the only comparable unsupervised baseline, and we could not reproduce their results. To solve this shortcoming, we propose an evaluation focusing on two directions.

Few-Shot & Zero-Shot Baselines First, we compare CITRUN with state-of-the-art few-shot and zero-shot NER models. These models are more supervised than CITRUN and thus expected to achieve better results than us. However, they allow us to contextualize the performance of unsupervised NER with more usual and standard low-resource approaches.

For few-shot models, we evaluate PromptNER [278] and MANNER [277] in a 10-shot setting. MANNER implements a two-step approach with BIO extraction and prototypical network-based entity typing. PromptNER uses prompt tuning with BERT. Support sets are selected with the widely used method of Hou et al. [302].

Regarding zero-shot approaches, we include the current state-of-the-art LLM-based UniNER [31], GoLLIE [23], and ChatIE (GPT-3.5) [151]. We also evaluate ChatIE with the open-weight Llama 3 8B[3]: ChatIE (Llama 3). UniNER and GoLLIE are LLMs fine-tuned on synthetically or manually labeled datasets, whereas ChatIE implements "raw" prompting. We also test GliNER L [66] and GNER [139], which respectively use an EncLM (DeBERTav3 [293]) and a full transformer (Flan-T5 [133, 134]), both fine-tuned on the same dataset as UniNER.

Unsupervised Baselines Creation As no unsupervised baseline is currently evaluable, we propose creating two baselines based on LLMs.

First, Zhou et al. [31] annotated the Pile-NER dataset by prompting GPT-3.5 without specifying entity types (see section 3.1 of their paper), thus in an unsupervised setting. They never evaluated this approach, and we include it to provide reference values of unsupervised GPT-3.5 prompting. We call this baseline UniNER Uns (GPT-3.5). The prompt they used is displayed in figure V.2. Additionally, we tried to replace GPT-3.5 with Llama 3 8B, but this smaller model could not respect the format specified in the prompt, resulting in null scores.

[3]Available at https://huggingface.co/meta-llama/Meta-Llama-3-8B.

1. Type Elicitation Prompt

   System Message: A virtual assistant answers questions from a user based on the provided text.

   Prompt: Given document: {***d***}. Please answer: What types of entities are included in this sentence? Answer with a JSON list like: [“entity type 1”, “entity type2”, ...].

2. Entity Extraction Prompt

   System Message: A virtual assistant answers questions from a user based on the provided text.

   Prompt: According to the document above, please output the entities of type “{e}” in the form of a JSON list like: [“entity 1”, “entity 2”, ...].

Figure V.3: Unsupervised adaptation of the prompting method of ChatIE [151]. ChatIE follows a multi-turn question-answering setup, with the first prompt employed to identify the entity types mentioned in the current document and subsequent questions to identify mentions for each elicitated entity type. This prompt is employed by ChatIE Uns (GPT 3.5) and ChatIE Uns (Llama 3).

Second, the dual-stage method that ChatIE [151] implements, with type elicitation and entity extraction, can be translated to work under an unsupervised setting. Initially, type elicitation necessitates the list of entity types $\mathcal{E}_T$, but we can reformulate it to remove this dependency. The prompts employed are displayed in figure V.3. We call this baseline ChatIE Uns (GPT-3.5). We could successfully replace GPT-3.5 with Llama 3 8B, and we call this approach ChatIE Uns (Llama 3). Finally, this baseline allows us to compare the performance between nearly identical models (ChatIE and ChatIE Uns) and observe the impact of not specifying entity types beforehand.

Can Zero-Shot Approaches Be Directly Translated to an Unsupervised Setting? The zero-shot and unsupervised settings are very similar, not needing annotated data in $\mathcal{D}_T$; the only difference is specifying entity types beforehand (zero-shot) or automatically discovering them (unsupervised). At first glance, the reader may think that zero-shot approaches can be easily translated to unsupervision. But the truth is more complex.

Fine-tuned approaches (EncLMs [66], full transformers [139] or LLMs [31]) all require the specification of an entity types schema, which is heavily employed during their training procedure. For instance, Ding et al. [139] experimentally observed that negative sampling (i.e., specifying entity types not mentioned in the current document) is necessary to attain state-of-the-art performances. Adapting these kinds of zero-shot approaches requires the redefinition of the training procedure, a nontrivial process beyond the scope of baseline evaluation.

Prompting of frozen LLMs [28, 151] is easier to adapt, as it necessitates only adjusting the prompt to remove the dependency on pre-specified entity types. That is precisely the models we have implemented, with a single-stage approach (UniNER Uns) and a multi-turn method

(ChatIE Uns).

V.4.2 Datasets

V.4.2.1 Target Domain $\mathcal{D}_T$

Specific domains are the primary use cases of an unsupervised NER. We focus on datasets that differ from $\mathcal{D}_S$ stylistically (types of text), semantically (topics), and/or from the entity type perspective (unseen entity types). As a result, we evaluate CITRUN on 13 domain-specific datasets:

- five CrossNER datasets [276] (*AI*, *Literature*, *Music*, *Politics*, and *Science*). They cover specific topics (scientific and literary) and unseen entity types.
- two MIT datasets [303] (*Movie* and *Restaurant*). They cover new styles of text (reviews and search engine queries), specific topics, and unseen entity types.
- *FabNER* [304] with physics and chemistry articles labeled with scientific entity types.
- *GENIA* [305] and *i2b2* [306] contain biomedical articles (taken from PubMed) annotated with biomedical entities.
- *GENTLE* [307] and *GUM* [308] cover unusual styles of text: e.g., dictionary entries, travel guides, legal notes, or poetry.
- *WNUT 17* [309] comprises social network posts.

These datasets cover a wide spectrum of types of text (encyclopedic, scientific, biomedical, social networks, customer reviews, dictionary entries, ...); domains (computer science, physics, chemistry, natural science, biomedical, literature, music, ...); and entity types (*algorithm*, *protein*, *cell type*, *poem*, *mechanical property*, *animal*, or *political party* among many others). It allows us to have a detailed picture of the quality and generalizability of CITRUN.

V.4.2.2 Source Domain $\mathcal{D}_S$

We propose to train CITRUN with two datasets: *CoNLL-2003* [39], and *Pile-NER* [31]. They represent two different ways to envision the unsupervised setting.

CoNLL represents the cross-domain perspective. It contains general-domain newspaper articles manually annotated with four entity types (*person*, *location*, *organization*, and *misc*). CoNLL is chosen to be distant from $\mathcal{D}_T$ stylistically, semantically, and from the entity type point of view. It allows us to evaluate the cross-domain capabilities of CITRUN.

Pile-NER represents the synthetic data perspective. It comprises 50,000 documents gathered from the Pile corpus [159] automatically annotated by GPT-3.5[4], resulting in 13,000 fine-grained

[4]As an aside, it is interesting to notice that Gao et al. employed "real" documents from the Pile corpus instead of generating them with GPT-3.5. They argue having diverse documents and wide coverage of domains with LLM-generated documents is difficult, resulting in lower performance.

entity types. The idea is that large and diverse $\mathcal{D}_T$ datasets benefit the generalizability and partially close the stylistic, semantic, or entity type gap between $\mathcal{D}_S$ and $\mathcal{D}_T$. Nonetheless, as the annotation process is automatic and does not involve human actions, it is not time-consuming or expensive. In fact, the latest few-shot and zero-shot models use large amounts of automatically annotated $\mathcal{D}_S$ data (e.g., UniNER, GliNER L, and GNER train on Pile-NER), and the results show the benefits of these automatically labeled corpora. Additionally, training CITRUN on Pile-NER is a way to ensure a fairer comparison with these baselines.

V.4.3 Metrics

V.4.3.1 Mention Detection

We use the binary F1 score, precision, and recall as defined in section III.5.1. A prediction is correct when the predicted boundaries are the same as those of a valid entity mention. Type is ignored for mention detection.

We notice that some recent LLM-based approaches [23, 31, 151] have changed the boundary check by a surface form check (i.e., checking that a predicted mention has the same text as a true mention)[5]. This modification is less precise than an exact boundary check and can be problematic when multiple mentions with the same surface form in the same document have different types (e.g., "French persons speak French"). In our evaluation, we have evaluated all baselines (and CITRUN) with the same boundary check metrics to ensure maximal fairness.

V.4.3.2 Entity Typing & End-to-End Named Entity Recognition

We cannot use the F1 score, as CITRUN predicts clusters, meaning there is no direct link between clusters and entity types. We employ the Adjusted Mutual Information (AMI) presented in section III.5.6 because most datasets have an unbalanced entity type distribution. To recall, the range of AMI is $[-1, 1]$, and higher values are better. AMI is adjusted for chance: random clustering produces a score close to 0.

V.4.4 Implementation Details

CITRUN follows a "train once, test anywhere" [310] methodology: it needs to be trained once on $\mathcal{D}_S$ and can be applied to multiple $\mathcal{D}_T$ datasets without further effort. Regarding hyperparameters, as CITRUN is unsupervised, we cannot use validation data to adjust them. We opt for standard hyperparameter values defined by Devlin et al. [6].

Entity Extraction We use DeBERTa v3 embeddings [293, 311], train the model for 4 epochs, using the Adam optimizer [312], a decreasing linear schedule without warmup, a learning rate of 2×10^{-5}, a batch size of 32, and dropout ($p = 0.1$) between the EncLM and the linear classifier.

[5] In fact, this change is not documented in their respective papers, but it is present in their source code.

Entity Typing We use BERT embeddings. We employ the simplest prompt possible, defined in equation (V.2), and train the model for 4 epochs, using the Adam optimizer [312], a decreasing linear schedule without warmup, a learning rate of 2×10^{-5}, a batch size of 128 as discussed in section V.3.2.3, and dropout ($p = 0.1$). For the brute force cluster estimation, we fix the upper bound K to 50 and increase it if $\hat{k}$ is close to K ($K = 100$ for GUM, $K = 100$ for CITRUN trained on CoNLL and tested on i2b2 and $K = 500$ for Pile-NER and i2b2).

Computational Resources Experiments were run on a single machine with 12 cores, 128 GB of RAM, and a GPU with 48 GB of VRAM. The required computational time is equivalent to BERT fine-tuning and depends on the size of the training dataset. With CoNLL, training usually last 50 min, and with Pile-NER, 5 h.

V.5 Results & Analysis

For CITRUN, each experiment is repeated with five random seeds, and we report the average value and the standard deviation.

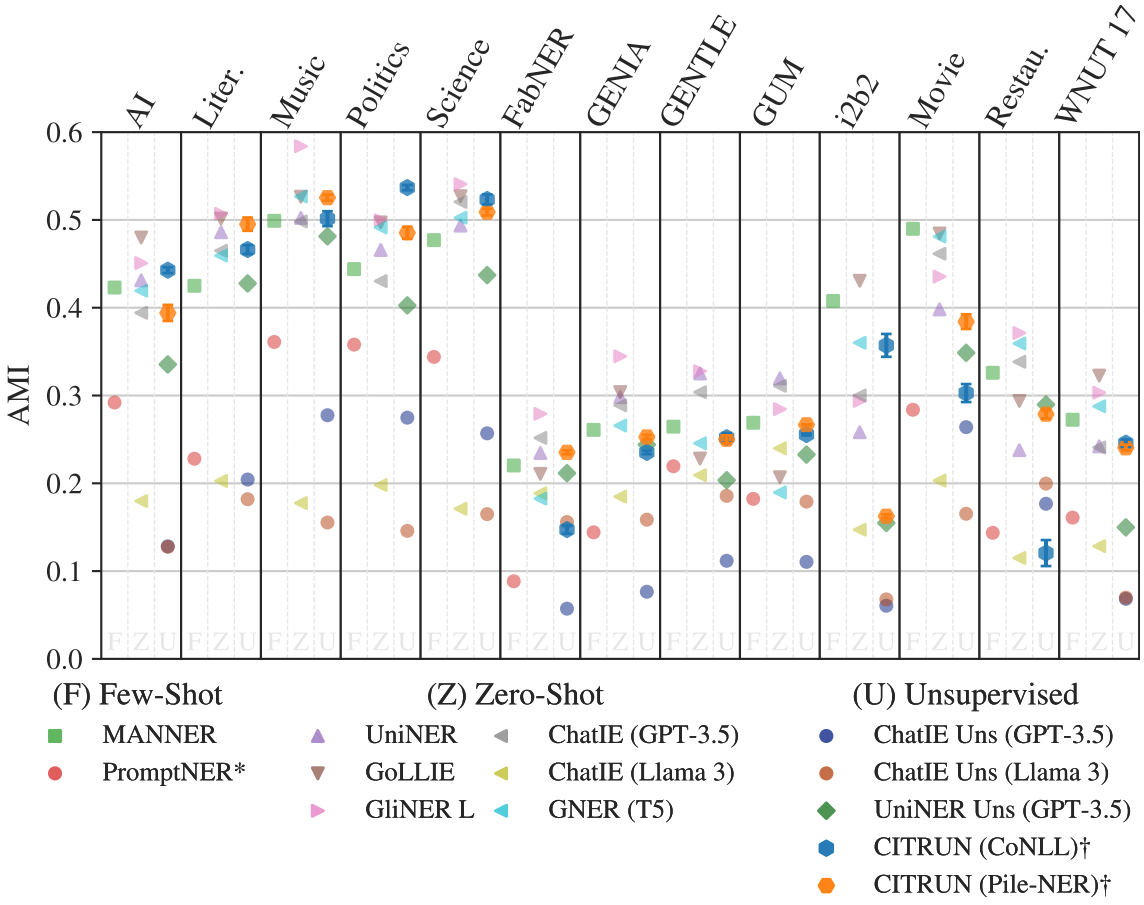
V.5.1 Comparison With the Baselines

The performances of CITRUN and the few-shot, zero-shot, and unsupervised baselines on the 13 $\mathcal{D}_T$ datasets are reported in figure V.4 and table V.1.

First, ● CITRUN (Pile-NER) outperforms ● UniNER Uns (GPT-3.5) with an average AMI gap of 4.3 % and beats it on 12 datasets. ● UniNER Uns (GPT-3.5) has a short advantage of 1 % in AMI for Restaurant. ● CITRUN (Pile-NER) surpasses ● ChatIE Uns (GPT-3.5) and ● ChatIE Uns (Llama 3) with an average AMI gap of 18 % and 19 %. ● CITRUN (CoNLL), trained on a much more distant $\mathcal{D}_T$ dataset, surpasses UniNER Uns (GPT-3.5), ChatIE Uns (GPT-3.5), and ChatIE Uns (Llama 3) with average AMI gaps of 3.6 %, 17 %, and 19 %. CITRUN, with its simple EncLM embeddings and architecture, performs significantly better than LLM-based unsupervised NERs.

Even compared to the more supervised zero-shot and few-shot models, CITRUN is not out of the picture. ● CITRUN (Pile-NER) performs significantly better than ◀ ChatIE (Llama 3) or ● PromptNER, and matches or surpasses the performances of ▲ UniNER on six datasets, ◀ ChatIE (GPT-3.5) on five datasets, ◀ GNER on five datasets, ▼ GoLLIE on four datasets, ■ MANNER on three datasets, and ▶ GliNER L on one dataset. Without accessing annotated data in $\mathcal{D}_T$ nor knowing the target entity types $\mathcal{E}_T$, CITRUN is competitive with the state-of-the-art zero-shot and few-shot approaches.

Additionally, the comparison between ChatIE (GPT-3.5) and ChatIE Uns (GPT-3.5) is interesting. As stated in section V.4.1, translating zero-shot baselines to unsupervised ones seems trivial at first glance. We insisted on the fact that fine-tuned baselines were not easily adaptable to work unsupervised, as their training is entirely based on the presence of an entity-type schema. For the frozen-LLM-based ones, we have proposed to modify ChatIE prompting to work fully unsupervised. We can see in table V.1 that ChatIE Uns performances are extremely low compared to ChatIE (zero-shot), the average gap is 21 %. The small modification



* We could not run PromptNER on i2b2 due to excessive RAM consumption.
† The standard deviation for CITRUN is displayed as a vertical bar.

Figure V.4: NER performances (AMI) of CITRUN, few-shot, zero-shot, and unsupervised baselines. CITRUN is less supervised than zero-shot and few-shot baselines and smaller than zero-shot and unsupervised baselines. Exact values can be seen in table V.1.

Table V.1: NER performances (AMI %) of CITRUN, few-shot, zero-shot, and unsupervised baselines. The best AMI for each $\mathcal{D}_T$ dataset and setting (few-shot, zero-shot, unsupervised) is in **bold**, and the best AMI for each $\mathcal{D}_T$ dataset is in green.

	AI	Liter.	Music	Politics	Science	FabNER	GENIA
Few-Shot							
MANNER	**42.3**	**42.5**	**49.9**	**44.4**	**47.7**	**22.0**	**26.1**
PromptNER	29.2	22.8	36.1	35.8	34.4	8.8	14.4
Zero-Shot							
UniNER	43.1	48.6	50.2	46.6	49.4	23.5	29.8
GoLLIE	**48.0**	50.2	52.6	49.7	52.7	21.1	30.4
GliNER L	45.1	**50.7**	**58.4**	**50.0**	**54.1**	**27.9**	**34.5**
ChatIE (GPT-3.5)	39.4	46.5	49.8	43.0	52.1	25.2	28.9
ChatIE (Llama 3)	18.0	20.3	17.8	19.8	17.1	18.9	18.5
GNER (T5)	41.9	45.9	52.7	49.1	50.2	18.3	26.6
Unsupervised							
UniNER Uns (GPT-3.5)	33.5	42.8	48.1	40.3	43.7	21.2	24.4
ChatIE Uns (GPT-3.5)	12.8	20.4	27.8	27.5	25.7	5.7	7.6
ChatIE Uns (Llama 3)	12.8	18.2	15.5	14.6	16.5	15.6	15.9
CITRUN (CoNLL)	**44.3(3)**	46.6(5)	50.1(9)	**53.7(3)**	**52.3(6)**	14.7(5)	23.5(2)
CITRUN (Pile-NER)	39.4(9)	**49.5(8)**	**52.5(3)**	48.5(7)	50.9(4)	**23.5(2)**	**25.3(3)**

	GENTLE	GUM	i2b2	Movie	Restau.	WNUT 17
Few-Shot						
MANNER	**26.5**	**26.9**	**40.8**	**49.0**	**32.6**	**27.2**
PromptNER	21.9	18.2	*	28.4	14.4	16.1
Zero-Shot						
UniNER	32.5	**32.0**	25.8	39.8	23.8	24.2
GoLLIE	22.8	20.7	**43.1**	**48.5**	29.4	**32.3**
GliNER L	**32.8**	28.4	29.4	43.6	**37.1**	30.3
ChatIE (GPT-3.5)	30.4	31.1	30.0	46.2	33.8	24.1
ChatIE (Llama 3)	20.9	24.0	14.7	20.3	11.5	12.8
GNER (T5)	24.6	19.0	36.0	48.1	35.9	28.8
Unsupervised						
UniNER Uns (GPT-3.5)	20.4	23.3	15.5	34.9	**29.0**	15.0
ChatIE Uns (GPT-3.5)	11.2	11.1	6.1	26.4	17.7	6.8
ChatIE Uns (Llama 3)	18.6	17.9	6.8	16.5	20.0	7.0
CITRUN (CoNLL)	**25.2(5)**	25.6(1)	**35.7(1.3)**	30.3(1.0)	12.1(1.5)	**24.6(4)**
CITRUN (Pile-NER)	25.0(4)	**26.7(1)**	16.2(2)	**38.4(8)**	27.9(5)	24.0(3)

The standard deviation of CITRUN is printed in parentheses.

* We could not run PromptNER on i2b2 due to excessive RAM consumption.

Table V.2: Number of parameters of CITRUN and few-shot, zero-shot, and unsupervised baselines. CITRUN is more than 60-70 times smaller than LLM-based NERs.

	Model	Backbone	# Parameters
Few-Shot	MANNER	BERT	110 M
	PromptNER	BERT	300 M (×2.7)
Zero-Shot	UniNER	Llama	7 B (×60)
	GoLLIE	Code-Llama	7 B (×60)
	GliNER L	DeBERTa v3	300 M (×2.7)
	ChatIE (GPT-3.5)	GPT 3.5	†
	ChatIE (Llama 3)	Llama 3	8 B (×70)
	GNER	Flan T5	275 M (×2.5)
Unsupervised	UniNER Uns (GPT-3.5)	GPT 3.5	†
	ChatIE Uns (GPT-3.5)	GPT 3.5	†
	ChatIE Uns (Llama 3)	Llama 3	8 B (×70)
	CITRUN	DeBERTa v3 / BERT	110 M*

† Although not disclosed, GPT-3.5 is expected to be larger than Llama 3.
* CITRUN uses two encoders with a total of 200 M parameters (90 M for DeBERTa v3 and 110 M for BERT). But at any given time, only one is loaded.

of removing the predefined list of entity types tremendously impacts performance. When looking closely, ChatIE Uns does not group together entities in coherent entity types and results instead in predicting overspecific entity types. For instance, ChatIE has identified 12,621 entity types (instead of 10) on the GUM dataset, such as *lantern festival*, *theme music*, *light show*, *laser light show*. As a side note, it is also a problem of UniNER Uns, at a lesser degree, though (see section V.5.5). This demonstrates that knowing entity types is a significant supervision signal (for zero-shot baselines), and removing it to achieve unsupervision is very challenging.

Finally, it is interesting to put the size of the compared baselines in perspective with the performances (see table V.2). CITRUN is the smallest model with its 110 M parameters (equally with MANNER), yet it competes with much larger LLM baselines that are one to two orders of magnitude bigger. In compute-constrained environments, CITRUN is a good alternative to larger models, especially LLMs. As an aside, contrary to the intuitive belief that LLMs perform particularly well in unsupervised and zero-shot settings, CITRUN (Pile-NER) and GliNER L (both with EncLM) have very competitive results on our 13 domain-specific $\mathcal{D}_T$ datasets.

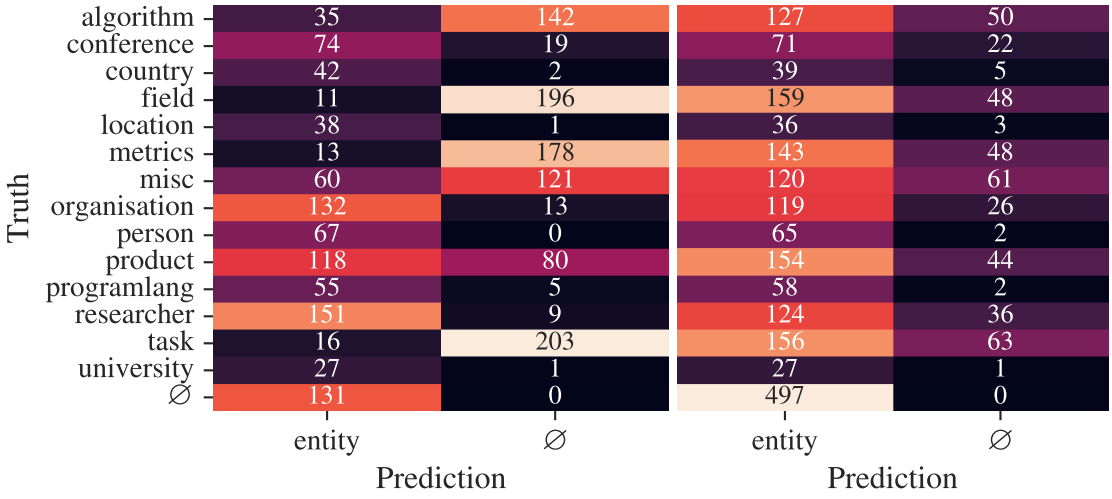
V.5.2 Cross-Domain Capabilities & Synthetic Annotations

The results in figure V.4 demonstrate that CITRUN works well with a distant $\mathcal{D}_T$ (CoNLL) and with synthetic data (Pile-NER), as they both lead to better performances than unsupervised baselines. ● CITRUN (Pile-NER) has a slight 0.7 % advantage in AMI compared to ● CITRUN (CoNLL).

However, when looking closely, CoNLL and Pile-NER build models with different behaviors

Table V.3: Comparison of precision (P) and recall (R) (in %) of CITRUN for MD between CoNLL and Pile-NER. The standard deviation is displayed in parentheses. Each $\mathcal{D}_T$ dataset's best precision and recall are in **bold**.

	$\mathcal{D}_S$	CoNLL		Pile-NER	
		P	R	P	R
$\mathcal{D}_T$	AI	**86.2(2)**	46.6(5)	74.1(6)	**77.5(5)**
	Liter.	**87.2(8)**	**80.9(4)**	85.5(3)	77.6(2)
	Music	84.1(4)	74.5(2)	**85.5(3)**	**82.8(4)**
	Politics	78.9(3)	**82.3(2)**	**82.1(5)**	81.2(3)
	Science	**82.8(5)**	67.6(9)	81.1(5)	**81.3(7)**
	FabNER	**52.0(7)**	5.5(3)	25.8(6)	**18.6(7)**
	GENIA	**46.5(1.2)**	27.1(1.4)	46.3(2)	**60.9(1.0)**
	GENTLE	32.0(1.2)	6.8(2)	**33.1(6)**	**20.6(3)**
	GUM	25.8(2)	6.4(1)	**28.6(1)**	**14.2(3)**
	i2b2	**22.1(2.2)**	26.8(1.4)	5.5(2)	**29.6(9)**
	Movie	**89.9(1.6)**	23.0(6)	71.8(1.2)	**46.0(9)**
	Restau.	**57.4(3.1)**	4.0(1.0)	51.8(1.0)	**32.6(9)**
	WNUT 17	**57.1(7)**	74.1(1.2)	41.1(4)	**76.6(4)**



(a) CITRUN trained on CoNLL, tested on AI.

(b) CITRUN trained on Pile-NER, tested on AI.

Figure V.5: Confusion matrices of CITRUN for MD tested on AI. The ∅ row shows the false positives, and the ∅ column shows the false negatives per entity type.

Table V.4: MD performances (F1 %) for different architectures trained on CoNLL and tested on five $\mathcal{D}_T$ datasets. The standard deviation is displayed in parentheses. We did not repeat the experiment for PURE and SpanProto as they were very slow to train. Each $\mathcal{D}_T$ dataset's best F1 score is in **bold**.

	AI	Liter.	Music	Politics	Science
BIO (CITRUN)	**60.5(4)**	**83.9(5)**	**79.0(2)**	**80.6(2)**	**74.4(7)**
PURE	39.8	37.1	33.8	32.4	35.7
SpanProto	54.1	62.9	59.6	68.7	59.7
WL-Coref	57.4(8)	68.3(1.7)	66.9(2.1)	72.1(1.3)	63.4(3.3)

(although similar performances). In table V.3, we display the precision and recall of CITRUN for mention detection. Overall, CITRUN tends to have more precision when trained on CoNLL and more recall when trained on Pile-NER. This is expected: the diversity of Pile-NER helps CITRUN detect entities better, while the human quality of annotations in CoNLL helps CITRUN be more precise. This observation is confirmed when we examine the confusion matrices in figure V.5. On one side, Pile-NER leads to better detections of domain-specific entity types (such as *algorithm*, *field*, *metrics*, or *task*), but we also see an increase in false positives (497 for Pile-NER vs. 151 for CoNLL). On the other side, CoNLL has a slightly better recall for *person*, *location*, or *organization*, which are precisely the entity types annotated in this dataset.

The question of higher false positives with Pile-NER is interesting. We manually checked them: from the 497 false positives, 53 % of them are correct mentions not annotated in AI, 42 % intersect with a true mention (boundary problem), and 5 % are wrong predictions. Overall, the boundary problem explains the false positive gap between CoNLL and Pile-NER, probably resulting from Pile-NER's imperfect annotations.

The 53 % correct entities not annotated in AI come from existing entity types (most missing mentions are acronyms, for instance, FPR = false positive rate) and new entity types (not in the 14 entity types annotated in AI). The fact that CITRUN identifies correct entity mentions of new entity types highlights its novelty detection capabilities. This behavior cannot be observed with the other zero-shot and few-shot baselines as they have a predefined set of entity types.

In conclusion, the cross-domain capabilities of CITRUN are highlighted by the good results of CITRUN (CoNLL) on the $\mathcal{D}_T$ datasets. Broadly speaking, the manual annotations of CoNLL bring precise results, and the diversity of Pile-NER provides better recall at the cost of precision. In a novelty detection or exploratory scenario, where recall is key, we advise the reader to use Pile-NER. Additionally, the analysis of the confusion matrices shows that CITRUN identifies mentions of novel entity types that are unknown beforehand.

V.5.3 BIO Sequence Labeling for Mention Detection

In section V.3.1, we propose to use a BIO extractor for MD, as we expect the simplicity of this architecture to bring better generalizability on new $\mathcal{D}_T$. In table V.4, we report the F1 score of different MD architectures, trained on CoNLL and tested on five $\mathcal{D}_T$ datasets. We evaluate the

Table V.5: AMI scores (in %) of CITRUN for ET on $\mathcal{D}_T$ datasets, without ER and with ER on CoNLL or Pile-NER. ET is evaluated using gold entity spans. The standard deviation is printed in parentheses. The best AMI for each $\mathcal{D}_T$ dataset is in **bold**.

	AI	Liter.	Music	Politics	Science	FabNER	GENIA
w/o ER	43.0(1.4)	40.1(6)	47.8(8)	56.0(8)	56.1(9)	18.6(3)	20.3(7)
ER on CoNLL	**56.8(1.4)**	56.3(1.1)	60.9(5)	65.4(3)	**66.7(3)**	**26.7(7)**	26.6(8)
ER on Pile-NER	54.2(7)	**63.1(8)**	**64.2(5)**	**66.0(1.1)**	66.0(9)	24.1(8)	**31.7(6)**

	GENTLE	GUM	i2b2	Movie	Restau.	WNUT 17
w/o ER	15.6(5)	19.7(2)	32.1(6)	21.8(5)	11.3(4)	22.5(3)
ER on CoNLL	21.5(7)	26.1(5)	47.9(6)	46.6(1.3)	35.8(1.4)	34.3(1.1)
ER on Pile-NER	**32.7(5)**	**37.0(2)**	**49.4(8)**	**52.1(8)**	**41.0(5)**	**41.1(8)**

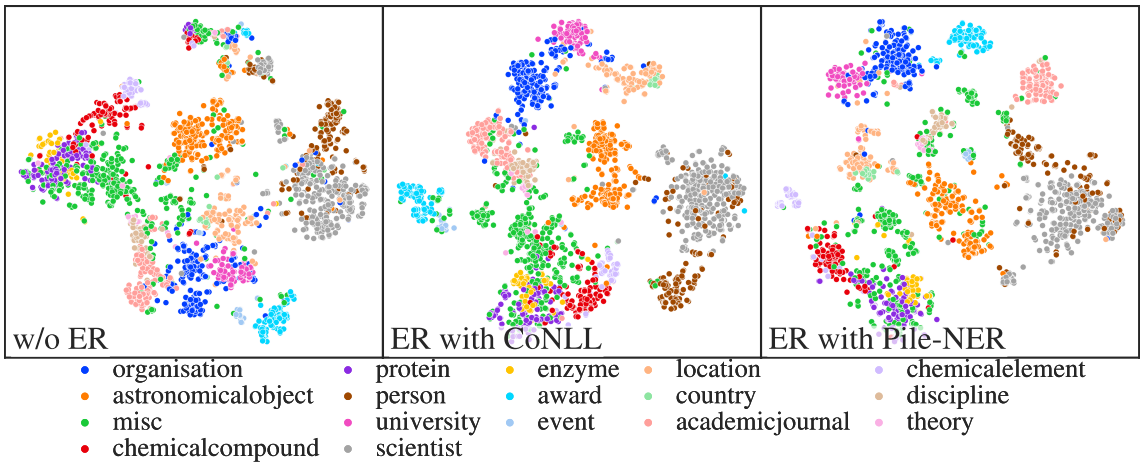
following architectures:

- BIO. It is the architecture implemented by CITRUN.
- PURE [113]. A span-based extractor that combines the start and end embeddings of a candidate span with a perceptron. It is the architecture we implemented for the PNA baseline of Linked-DocRED (see section III.6.1).
- SpanProto [116]. A span-based extractor that uses bilinear neurons to combine start and end embeddings of a candidate span (which provides faster predictions compared to PURE).
- WL-Coref [58]. A span-based extractor that identifies the "head" of the mention and recomposes its boundaries using a convolutional network. This model tackles the quadratic complexity problem of traditional span-based extractors.

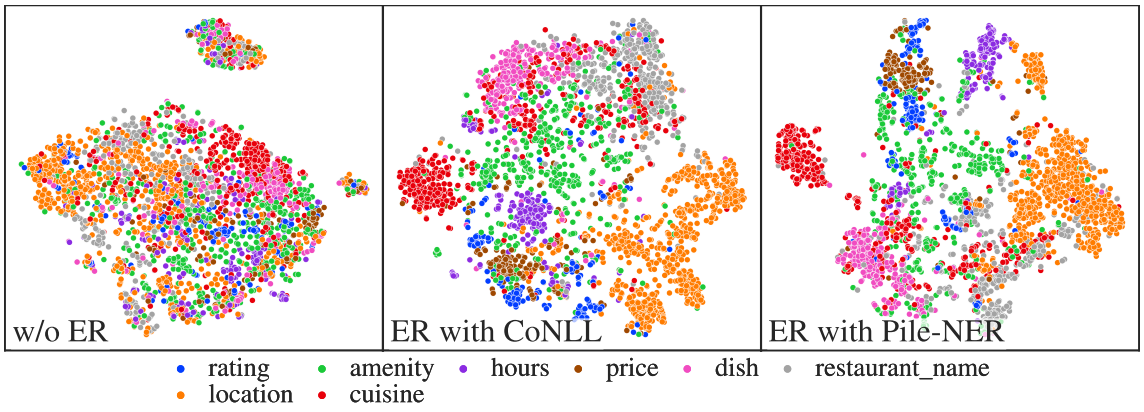
In a fully supervised setting, PURE, SpanProto, and WL-Coref are shown to be slightly better than BIO sequence labeling [58, 113, 116]. However, in our unsupervised cross-domain setting, BIO performs significantly better than span-based extractors, with an average gap of 40 % with PURE, 15 % with SpanProto, and 10 % with WL-Coref, while being faster to train. We believe that the simplicity of our BIO architecture reduces overfitting and benefits generalizability on new domains. As an aside, this observation was made by Fang et al. [277], who also use BIO sequence labeling for their few-shot MANNER model.

V.5.4 Impact of the Embedding Refinement

An important component of CITRUN is embedding refinement (ER), which aims to improve EncLM representations for entity clustering using contrastive learning. In table V.5, we compare CITRUN entity typing performance without ER and with ER trained on CoNLL or Pile-NER.



(a) CITRUN tested on Science.



(b) CITRUN tested on Restaurant.

Figure V.6: Two-dimensional t-SNE visualizations of the entity embeddings of CITRUN. For each subfigure from left to right: 1. without ER, 2. ER with CoNLL, and 3. ER with Pile-NER.

We use the gold entity spans from $\mathcal{D}_T$ (no MD) to assess only the effect of ER. This is why the AMI scores are higher than in table V.1.

We see that ER has a significant positive impact with CoNLL and Pile-NER on each of the 13 $\mathcal{D}_T$ datasets, with an average AMI gain of 12.8 % for CoNLL and 16.7 % for Pile-NER compared to CITRUN without ER. The gain is particularly impressive for datasets that are difficult for raw BERT embeddings, such as GENTLE, GUM, i2b2, Movie, Restaurant, or WNUT 17. Pile-NER's better performances can be explained by its diversity of entity types (13,000 entity types), which helps to fine-tune entity embeddings more precisely. Nevertheless, CoNLL achieves honorable performances despite only having four entity types. This validates the hypothesis that refining entity embeddings on $\mathcal{D}_S$ with contrastive learning benefits also distant $\mathcal{D}_T$.

To give a more visual representation of the effects of embedding refinement, we display in figure V.6 two-dimensional t-SNE [313] representations of the entity embeddings of the Science and Restaurant datasets. The entities of Science are already well isolated without ER (see table V.5). Still, we can notice several improvements: better separation of *discipline*, *organization*, and *academicjournal* (CoNLL and Pile-NER); better separation of *chemicalelement* and *chemicalcompound* (Pile-NER); and the multi-type cluster at the top of the w/o ER figure has disappeared. The effects of ER are more visible with the difficult Restaurant dataset: without ER, ET cannot discriminate any entity type, and we see huge improvements with ER on CoNLL or Pile-NER. In particular, it is interesting to see that ER with CoNLL leads to a relatively good separation of *cuisine*, *hours*, or *price*, even though CoNLL does not contain such entities. The effects are more complete and more visible with Pile-NER.

In conclusion, ER significantly improves ET performance with CoNLL and Pile-NER. The best results are achieved with Pile-NER due to its diversity in entity types. ER works well with the distant $\mathcal{D}_T$ dataset CoNLL, with noticeable improvements on unseen entity types. It also shows that ER is beneficial even with a labeled dataset with a narrow set of entity types (4 for CoNLL).

V.5.5 Estimation of the Number of Clusters $\hat{k}$

As we do not have any information about entity types (contrary to zero-shot approaches), CITRUN has to infer entity types and their number. In this part, we only consider the brute force cluster estimation. In table V.6, we display for each $\mathcal{D}_T$ dataset its true number of entity types k, the estimated number of clusters $\hat{k}$, the corresponding AMI score with $\hat{k}$ (similar to figure V.4), and AMI score with the ideal k.

Overall, CITRUN tends to overestimate the number of entity types; this effect is more pronounced with Pile-NER than with CoNLL. However, compared to UniNER Uns (GPT-3.5), CITRUN provides estimations that are much closer to the truth. Regarding Pile-NER, this overestimation behavior can be linked to its fine-grained entity types[6]. We can see this tendency in the visualization of Science in figure V.6, where the *misc* class is divided into multiple small

[6]As a reminder, Pile-NER was annotated using UniNER Uns (GPT-3.5). De facto, Pile-NER exhibits the same fine-grained entity type weakness as UniNER Uns (GPT-3.5). Fortunately, CITRUN partially mitigates this issue with a more reasonable estimation of the number of clusters, as shown in table V.6.

Table V.6: Estimation of the number of clusters $\hat{k}$ by CITRUN using the brute-force approach and AMI scores (in %) for NER with true k and estimated $\hat{k}$. The standard deviation is printed in parentheses. k and $\hat{k}$ are displayed in green, and the best AMI score for each $\mathcal{D}_T$ dataset is in **bold**. We also include the number of entity types found by UniNER Uns (GPT-3.5).

	AI	Liter.	Music	Politics	Science	FabNER	GENIA
k	14	12	13	9	17	12	5
CITRUN (CoNLL)							
$\hat{k}$	10	12(1)	20(2)	23(2)	17(2)	8(2)	20(1)
AMI $\hat{k}$	44.3(3)	**46.6(5)**	50.1(9)	53.7(3)	52.3(6)	14.7(5)	23.5(2)
AMI k	**44.5(3)**	46.4(4)	**51.0(3)**	**56.5(1.9)**	**52.9(6)**	**14.8(5)**	**26.4(6)**
CITRUN (Pile-NER)							
$\hat{k}$	18(1)	16(1)	26(1)	32(2)	29	32(3)	35
AMI $\hat{k}$	**39.4(9)**	49.5(8)	52.5(3)	**48.5(7)**	50.9(4)	23.5(2)	25.3(3)
AMI k	39.2(7)	**50.2(6)**	**54.3(4)**	47.8(6)	**51.7(5)**	**25.2(3)**	**29.0(5)**
UniNER Uns (GPT-3.5)							
$\hat{k}$	155	92	115	103	195	292	319

	GENTLE	GUM	i2b2	Movie	Restau.	WNUT 17
k	10	11	23	12	9	6
CITRUN (CoNLL)						
$\hat{k}$	8	35(1)	50(5)	8	4	8(1)
AMI $\hat{k}$	**25.2(5)**	25.6(1)	35.7(1.3)	**30.3(1.0)**	**12.1(1.5)**	**24.6(4)**
AMI k	25.1(7)	**27.0**	**38.7(9)**	29.8(1.0)	11.7(1.6)	**24.6(3)**
CITRUN (Pile-NER)						
$\hat{k}$	22(1)	59	197(4)	26	14(2)	16(1)
AMI $\hat{k}$	25.0(4)	26.7(1)	16.2(2)	38.4(8)	27.9(5)	24.0(3)
AMI k	**26.0(3)**	**28.8(1)**	**23.1(2)**	**38.7(9)**	**28.2(5)**	**25.1(6)**
UniNER Uns (GPT-3.5)						
$\hat{k}$	250	830	1,033	176	117	266

Table V.7: Estimation of the number of clusters $\hat{k}$ with brute force or ternary search and AMI scores (in %) for NER with true k and estimated $\hat{k}$, when CITRUN is trained on Pile-NER. The standard deviation is printed in parentheses. k and $\hat{k}$ are displayed in green, and the best AMI score for each $\mathcal{D}_T$ and $\mathcal{D}_T$ dataset is in **bold**.

	AI	Liter.	Music	Politics	Science	FabNER	GENIA
k	14	12	13	9	17	12	5
brute $\hat{k}$	18(1)	16(1)	26(1)	32(2)	29	32(3)	35
ternary $\hat{k}$	19(1)	18(1)	25(2)	32(3)	28(3)	32(4)	34(2)
AMI	39.2(7)	**50.2(6)**	**54.3(4)**	47.8(6)	**51.7(5)**	**25.2(3)**	**29.0(5)**
brute AMI	**39.4(9)**	49.5(8)	52.5(3)	48.5(7)	50.9(4)	23.5(2)	25.3(3)
ternary AMI	**39.4(7)**	49.2(6)	52.1(3)	**49.1(9)**	50.6(7)	23.4(2)	25.3(2)

	GENTLE	GUM	i2b2	Movie	Restau.	WNUT 17
k	10	11	23	12	9	6
brute $\hat{k}$	22(1)	59	197(4)	26	14(2)	16(1)
ternary $\hat{k}$	23(2)	65(5)	198(4)	24(2)	15(1)	18(1)
AMI	**26.0(3)**	**28.8(1)**	**23.1(2)**	**38.7(9)**	**28.2(5)**	**25.1(6)**
brute AMI	25.0(4)	26.7(1)	16.2(2)	38.4(8)	27.9(5)	24.0(3)
ternary AMI	25.0(5)	26.5(2)	16.2(2)	38.7(6)	28.0(3)	24.1(4)

clusters (compared to CoNLL).

AMI scores with the ideal k are close to AMI with $\hat{k}$ (AMI gap of 0.8 % for CoNLL and 1.5 % for Pile-NER on average), meaning that the clusterings are relatively similar from a qualitative point of view even with $\hat{k} \gg k$[7]. The long-tail distribution of the cluster membership explains this. If we take the second confusion matrix of figure V.8, a minority of clusters contains most entities, and the rest contain few specific entities. In fact, the 17 last clusters represent false positives[8] and members of the *misc* class (by definition, composed of multiple entity types). It explains why, even with this number of clusters, the performances do not plummet because the supplementary clusters model essentially false positives and composite classes.

V.5.6 Faster Estimation of the Number of Clusters $\hat{k}$

Up to this section, we used the brute force algorithm to estimate the number of clusters $\hat{k}$. The computational time is acceptable for the small datasets, but for the biggest $\mathcal{D}_T$ datasets (e.g., i2b2 or GUM), it can take up to hours (see table V.8), representing, in fact, the major part of the run. For instance, the cluster estimation lasts 13.6 h on average for $\mathcal{D}_S$ = Pile-NER and $\mathcal{D}_T$ = i2b2. This motivates the ternary search algorithm we presented in section V.3.2.2.

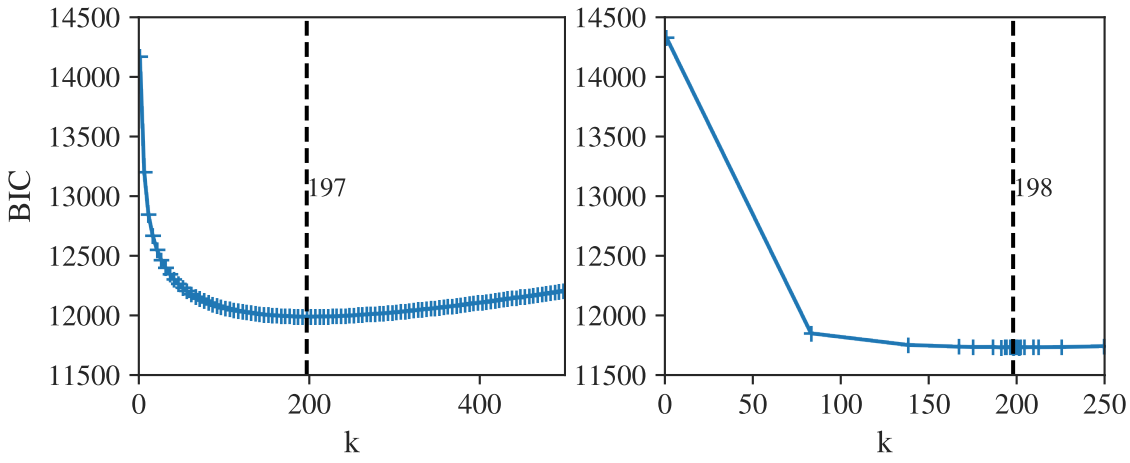
In table V.7, we display the comparison of the estimation of $\hat{k}$ between the brute force

[7]Except CITRUN (Pile-NER) tested on i2b2. I2b2 is a hard $\mathcal{D}_T$ dataset, even for the baselines (figure V.4).
[8]That can be correct entities, as we have seen in section V.5.2.

Table V.8: Execution time (in s) of the cluster estimation using the brute force or ternary search algorithms when CITRUN is trained on Pile-NER.

	AI	Liter.	Music	Politics	Science	FabNER	GENIA
brute AMI	88(2)	98(9)	191(23)	236(12)	164(15)	505(29)	390(36)
ternary AMI	**48(1)**	**52(1)**	**69(1)**	**89(3)**	**69**	**189(4)**	**142(5)**
	(÷1.8)	(÷1.9)	(÷2.8)	(÷2.6)	(÷2.4)	(÷2.7)	(÷2.7)

	GENTLE	GUM	i2b2	Movie	Restau.	WNUT 17
brute AMI	115(10)	1,823(157)	49,039(214)	276(28)	138(14)	99(11)
ternary AMI	**65(1)**	**906(28)**	**2,440(173)**	**117(2)**	**74**	**57(1)**
	(÷1.8)	(÷2.0)	(÷20.1)	(÷2.3)	(÷1.9)	(÷1.7)



(a) ET with brute force algorithm.

(b) ET with ternary search algorithm.

Figure V.7: BIC curves computed to estimate the number of clusters $\hat{k}$, when CITRUN is trained on Pile-NER and tested on i2b2. Each cross represents a computed clustering. With the brute force algorithm, 500 clusterings were calculated, and with ternary search, only 21.

Table V.9: MD performances (F1 %) of CITRUN trained on Pile-NER, using various EncLM embeddings. The standard deviation is printed in parentheses. The best F1 for each $\mathcal{D}_T$ dataset is in **bold**. The last column displays the average F1 across the 13 $\mathcal{D}_T$ datasets.

	AI	Liter.	Music	Politics	Science	FabNER	GENIA
BERT	73.7(5)	76.4(2)	80.3(2)	79.7(5)	78.4(3)	20.0(6)	50.7(3)
RoBERTa	74.3(5)	79.5(3)	81.9(4)	80.5(2)	78.8(3)	20.5(5)	51.2(5)
ERNIE	73.4(2)	76.0(4)	80.7(2)	80.1(4)	78.0(4)	20.6(2)	51.2(5)
ELECTRA	73.9(4)	76.3(3)	81.3(2)	79.6(4)	79.2(2)	20.5(3)	51.4(3)
DeBERTa v3	**75.6(5)**	**81.4(4)**	**84.6(3)**	**81.6(3)**	**80.9(5)**	**21.2(6)**	**52.2(4)**

	GENTLE	GUM	i2b2	Movie	Restau.	WNUT 17	Average
BERT	23.3(3)	**19.0(1)**	9.2(3)	56.9(7)	38.4(6)	47.6(6)	50.3
RoBERTa	23.9(7)	18.9(4)	**9.6(4)**	52.2(1.9)	39.8(6)	**54.2(8)**	51.2
ERNIE	22.6(5)	**19.0(2)**	9.4(2)	57.9(5)	40.0(7)	48.2(5)	50.5
ELECTRA	23.1(6)	18.2(3)	9.5(3)	**59.6(3)**	**41.5(6)**	48.5(6)	51.0
DeBERTa v3	**25.2(3)**	18.9(3)	9.5(1)	56.1(1.1)	39.8(8)	53.4(5)	**52.4**

algorithm and the ternary search, and the corresponding NER AMI scores; and in table V.8 we display the corresponding execution time. We see that the estimation of $\hat{k}$ with ternary search equals the brute force algorithm or is in the standard deviation range. This results in ternary search AMI scores virtually identical to brute force scores.

More interesting is the gain in terms of computational time. As displayed in table V.8, the ternary search is 1.7 to 2.7 quicker to run compared to the brute force algorithm, even on the smallest datasets. The gain is particularly impressive for the large i2b2 dataset with its large $\hat{k}$ where the gain is twenty-fold. Initially, the runs lasted 13.6 h, and with the ternary search, they are reduced to 41 min. The computational gain is less important for smaller sets of entity types (although still very significative) because of the slight rugosity of the BIC curve. This rugosity forces us to compute multiple clusterings sequentially once $k_{max} - k_{min} \leq 5$.

The case of the i2b2 dataset is especially interesting. In figure V.7, ternary search quickly converges to the minimum value without evaluating every possible $\hat{k}$. In particular, the range $[0, 140]$ clusters is eliminated in two steps (5 min), whereas brute force needs 2 h to evaluate the same interval. Ternary search finds $\hat{k}$ after 21 clusterings, compared to the 500 needed for the brute-force algorithm (24 times less).

In conclusion, the computational gain of ternary search is particularly important with large $\mathcal{D}_T$ datasets with many different entity types. It is also relevant for smaller datasets, bringing a two-fold decrease in calculation time. Empirically, we find no difference in the estimation of $\hat{k}$ and AMI scores between brute force and ternary search.

Table V.10: ET performances (AMI %) of CITRUN trained on Pile-NER, using various EncLM embeddings. ET is evaluated using gold entity spans. The standard deviation is printed in parentheses. The best AMI for each $\mathcal{D}_T$ dataset is in **bold**. The last column displays the average AMI across the 13 $\mathcal{D}_T$ datasets. The number of clusters is estimated using the ternary search algorithm, which explains why AMI scores are not identical to table V.5 (brute force). They are nevertheless in the range of standard deviation.

	AI	Liter.	Music	Politics	Science	FabNER	GENIA
BERT	**54.3(3)**	**64.1(1.0)**	64.4(5)	**66.2(1.3)**	**65.9(4)**	24.5(5)	**32.1(5)**
RoBERTa	53.7(8)	63.7(1.1)	**64.7(7)**	63.3(1.7)	65.7(1.0)	**25.0(7)**	28.1(4)
ERNIE	54.2(7)	62.7(1.4)	64.4(9)	63.0(1.2)	**65.9(8)**	24.7(4)	30.2(5)
ELECTRA	53.6(2)	57.9(2.0)	61.1(1.2)	56.7(1.5)	62.7(1.5)	22.7(7)	26.4(2.6)
DeBERTa v3	53.0(1.0)	59.0(1.1)	61.6(7)	58.5(1.3)	62.9(8)	**25.0(2)**	25.4(4)

	GENTLE	GUM	i2b2	Movie	Restau.	WNUT 17	Average
BERT	33.8(7)	35.7(1)	50.2(5)	52.0(3)	40.7(8)	40.9(1.2)	48.1
RoBERTa	**34.0(5)**	**36.2(3)**	**51.2(7)**	47.4(6)	47.2(6)	44.1(5)	48.0
ERNIE	33.8(6)	35.6(3)	48.5(3)	**54.2(8)**	47.3(1.0)	44.2(3)	**48.4**
ELECTRA	32.8(5)	34.1(1.2)	48.5(1.1)	51.8(9)	45.7(1.4)	41.8(6)	45.8
DeBERTa v3	33.4(3)	34.3(1)	50.8(5)	48.7(6)	**47.6(9)**	**46.7(6)**	46.7

V.5.7 Impact of the EncLM Embeddings

With CITRUN, we primarily utilize DeBERTa v3 [293, 311] for MD and BERT [6] for ET. In this section, we evaluate the performances of other popular EncLM such as RoBERTa [256], ERNIE [120], or ELECTRA [314].

In table V.9, we display the MD performances of various EncLMs when CITRUN is trained on Pile-NER, and in table V.10, we show the ET performances of the same EncLMs (also on Pile-NER). Broadly speaking, CITRUN works relatively well, regardless of the EncLM used as a backbone. Interestingly, the "older" model, BERT, is not out of the picture and performs similarly to more recent alternatives.

For MD, we see an advantage of DeBERTa v3 over the other approaches, with an average gap of 1.2 % with the second-best model RoBERTa. We link these better performances to the richer and broader pre-training dataset compared to the other EncLM. BERT achieves the worst performances. This explains why we have chosen DeBERTa v3 as the backbone for MD.

The performances are closer for ET, with BERT, RoBERTa, and ERNIE nearly indistinguishable (especially given the standard deviation). ELECTRA and DeBERTa v3 have lower AMI scores. The behavior of DeBERTa v3 is surprising, as it is generally recognized as the best-performing EncLM currently available. The performances of DeBERTa v3 are even worse without ER (not shown), achieving half of those of BERT without ER. The same conclusion can be drawn with ELECTRA. DeBERTa v3 and ELECTRA seem to have a less entity-type-oriented embedding space than BERT. As a result, we have chosen BERT embeddings for CITRUN.

ERNIE and RoBERTa would have also been valid choices.

V.5.8 Qualitative Analysis

We want to finish this analysis by giving a qualitative overview of the performances of CITRUN. In figure V.8, we display three confusion matrices of CITRUN trained with different $\mathcal{D}_S$ datasets and tested on different $\mathcal{D}_T$.

The three confusion matrices show a relatively clear diagonal, meaning that CITRUN correctly identifies most entity types. It is an impressive result: without annotated data in $\mathcal{D}_T$ nor any information on entity types or their count, CITRUN detects and structures entities in a scheme similar to the ground truth. The confusion matrices are, in fact, very similar to those of PromptORE (see figure IV.3), demonstrating that our prompt-based typing method works well to identify relation and entity types.

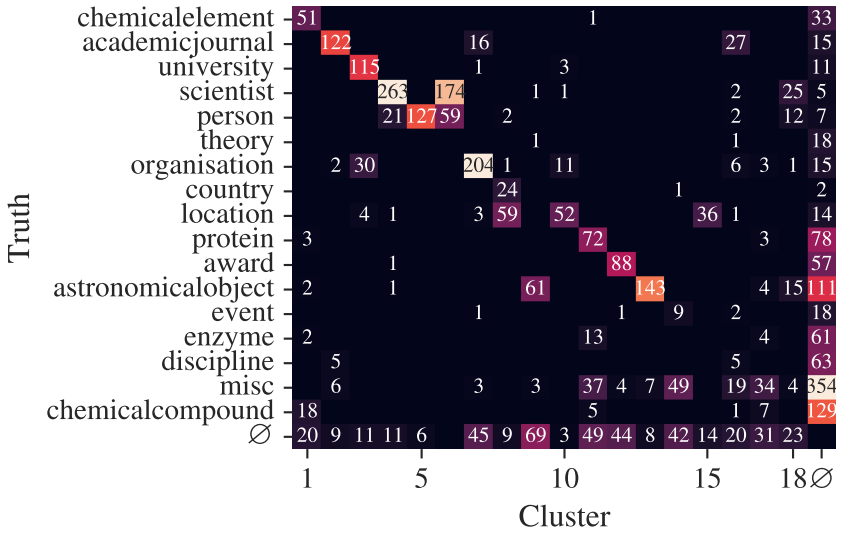
It is interesting to look at the confusions made by CITRUN. CITRUN merges *country* and *location* (Science and AI); *person* and *scientist*/*researcher* (Science and AI); *enzyme* and *protein* (Pile-NER Science); *task*, *product*, *field*, *algorithm* (AI); or *conference*, *university*, *organization* (AI). CITRUN confuses semantically close entity types, which is a reassuring behavior. It is also a constraint linked to unsupervised NER. As we do not provide the list of entity types, CITRUN organizes entities in a semantically coherent scheme that is a valid typing scheme but not exactly the dataset annotation schema.

Finally, CITRUN organizes false positives and *misc* entities, a composite of multiple underlying types. It explains why CITRUN tends to overestimate the true number of entity types.

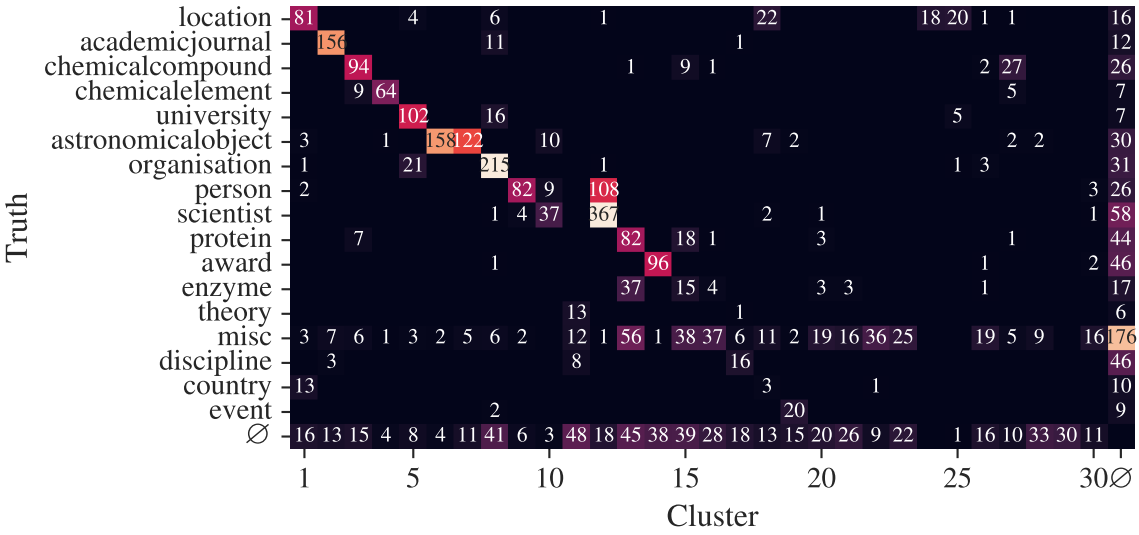
In conclusion, CITRUN organizes entities in a coherent typing scheme that is close to the true entity types. This analysis also highlights CITRUN's exploratory abilities. It can identify and organize entities into meaningful groups without labeled data in $\mathcal{D}_T$. CITRUN efficiently processes unannotated documents to uncover primary entities and their types, setting the stage for further refinement through more supervised methods.

V.6 Conclusion

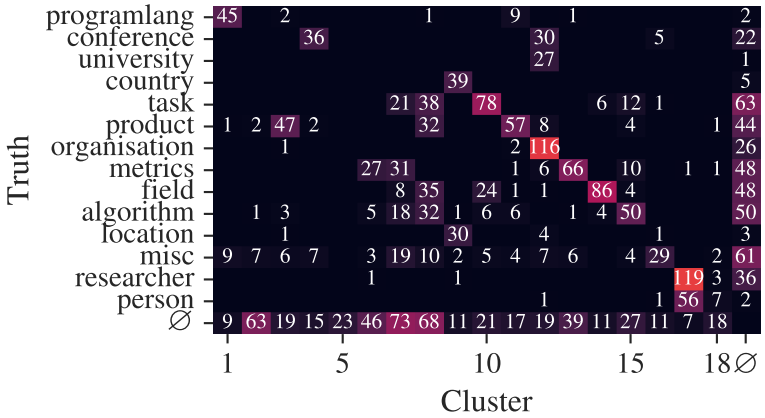
In this chapter, we presented CITRUN, our unsupervised and open-world NER model that transfers knowledge from $\mathcal{D}_S$ to $\mathcal{D}_T$ without supervision. The literature review showed that unsupervised NER lags while significant progress has been made towards lower resource NER (in particular, zero-shot NER). Most existing "unsupervised" models are not entirely unsupervised, as they rely on some forms of supervision or, similarly to unsupervised RE, on hyperparameters complex to adjust without in-domain validation data. CITRUN is proposed to be the first NER compatible with an utterly unsupervised scenario, to provide a strong baseline, and to stimulate further research. CITRUN is built upon the success of PromptORE for the entity typing part, keeping the model voluntarily simple. We have nevertheless refined it in several aspects. We proposed to apply contrastive learning using labeled data from $\mathcal{D}_S$ to elicit entity types better. We also implemented a faster algorithm to estimate the number of clusters $\hat{k}$.



(a) CITRUN trained on CoNLL, tested on Science.



(b) CITRUN trained on Pile-NER, tested on Science.



(c) CITRUN trained on Pile-NER, tested on AI.

Figure V.8: Confusion matrices of CITRUN for NER tested on various $\mathcal{D}_T$ datasets. Columns and rows were reordered using the algorithm described in appendix A. The ∅ row shows the false positives, and the ∅ column shows the false negatives.

Tests on 13 domain-specific datasets demonstrate that CITRUN outperforms LLM-based unsupervised NERs and is competitive with state-of-the-art zero-shot NER models without requiring prior knowledge of $\mathcal{D}_T$. This is a truly impressive result, given that the simple EncLM embeddings of CITRUN compete with much larger LLMs. We believe an essential point for CITRUN's success is its architectural simplicity and parameter efficiency, which achieve state-of-the-art results. This makes it easier to deploy in a realistic and constrained production environment.

Ablation studies show that ER brings significant performance gains and works well even with a distant $\mathcal{D}_T$ dataset. Ternary search shortens the computational time needed to estimate the number of clusters considerably (two times in general and up to twenty times faster on the largest dataset). Qualitative results demonstrate CITRUN's exploratory capabilities and ability to organize entities in semantically coherent clusters close to actual entity types.

VI Conclusion

Contents

We now arrive at the concluding chapter of this thesis. We started this manuscript by observing the exponential evolution of the quantity of information, which is increasingly at the economic core of our society. Most of this information is unstructured and remains underexplored. In particular, domain-specific unstructured document collections, such as biomedical literature, scientific articles, economic reports, etc., are goldmines but are just starting to be exploited due to the complexity of extracting reliable information. In chapter II, we summarized the impressive advances in information extraction. We also highlighted the current shortcomings:

- End-to-end IE models encompassing named entity recognition, coreference resolution, entity linking, and relation extraction are lacking.
- The more realistic document-level IE setting is underexplored and brings unique challenges, such as information scarcity and long context handling.
- Most IE approaches assume a closed world, where entity and relation types are predetermined, whereas the harness of unstructured documents requires open-world models.
- Annotated data is not always available, particularly in specific domains that are the most interesting to explore.

In chapter III, we explored the feasibility of end-to-end document-level IE by creating the first large-scale document-level IE dataset, proposing a complete set of evaluation metrics, and

evaluating baseline models on the dataset. In chapter IV, we proposed a novel architecture to tackle unsupervised relation extraction (both ensuring open-world capabilities and minimizing the quantity of annotated data required). This architecture was then generalized and improved for open-world named entity recognition in chapter V.

VI.1 Summary of the Contributions

VI.1.1 Dataset and Metrics to Evaluate Information Extraction Models

A prerequisite to developing end-to-end IE is a large, diverse, and document-level dataset, manually labeled for NER, CR, EL, and RE, to enable the training and evaluation of document-level IE models. Such a dataset was nonexistent. Therefore, our first contribution was to create Linked-DocRED.

To do that, we complemented the existing and widely used DocRED dataset with entity linking annotations. To minimize the annotation effort while maintaining human quality, we implemented a semi-automatic multi-step entity linking process. First, observing that DocRED documents are sourced from Wikipedia, we aligned Wikipedia article wikilinks with DocRED entities using the Needleman-Wunsch algorithm. This provided human-quality disambiguations (as human contributors edit wikilinks) at a low cost. Then, entity linking was complemented with common knowledge and contextual information, and the last undisambiguated entities were manually labeled. This process also allowed us to correct some entity and coreference errors of DocRED. Linked-DocRED contains more than 95,000 disambiguated entities, 38,000 coreferences, and 50,000 relations.

A second point missing from the state of the art was evaluation metrics. Indeed, the usual mention-level metrics fall short when long documents with large coreferent clusters are considered. These metrics are "hard", meaning a coreference cluster is considered entirely wrong if it lacks a single coreference (or contains a supplementary one). This leads to poorly discriminative metrics [21]. We proposed complementing the "soft" metrics defined by Zaporojets et al. [21] for the entity linking task and generalizing them for open-world IE replacing the notion of F1 score by the clustering V-measure, B^3, ARI, and AMI scores.

The evaluation of a solid end-to-end IE baseline demonstrated promising results, notably outperforming more complex multitask learning and integrated baselines. It also highlighted the challenges of cascading errors (imperfect NER and CR substantially impact RE and EL performances) and the complexity of handling long documents. Above all, our fully supervised and closed-world baseline achieved F1 scores that are far from perfect ($50\,\% - 60\,\%$), which raised the question of the current feasibility of an open-world and low-resource document-level IE.

VI.1.2 Towards Unsupervised Open-World Relation Extraction

Considering the previous uncertainty and question about the practicability of open-world and low-resource IE, we decided to first focus on a subset of IE, relation extraction, to develop a novel open and unsupervised method.

We noticed previous approaches were tested on simple datasets (general domain, few relation types), and critical parts of their implementation (unsupervised hyperparameter tuning) were unclear. In particular, the then state-of-the-art approach, SelfORE [197], was observed to be highly sensible to hyperparameter values that needed to be adjusted precisely for each test dataset, which is difficult to do in a "true" unsupervised setting.

As a result, we proposed PromptORE, which relies on EncLM prompting[1] to generate relation-type embeddings that are then clustered to identify groups of instances that express the same relation type. We were particularly vigilant in reducing the number of hyperparameters and providing unambiguous methods to adjust them unsupervised. Using prompting allowed us to avoid fine-tuning PromptORE and thus remove all training-related hyperparameters. We also proposed estimating the number of clusters (relation types) using the simple elbow rule method.

Experimental results showed that PromptORE outperformed previously state-of-the-art methods by a large margin (18 % – 20 % in B^3, V-measure, and ARI), while being simpler and not hyperparameter dependant. Qualitative analysis of PromptORE's confusion matrices also demonstrated that it organizes relation instances in clusters that follow a semantically coherent scheme close to the true relation types.

VI.1.3 Generalization to Open-World Named Entity Recognition

Encouraged by the successes of PromptORE, we then explored a second major task of IE, named entity recognition, under the same unsupervised and open-world setting.

Similarly to the previous contribution, we observed weaknesses in the experimental setup of state-of-the-art unsupervised baselines. The majority of them were not open-world [35, 36, 37, 38], as they required supervision for each entity type. On the other side of the spectrum, zero-shot approaches that have seen rapid progress were low-resource but assumed a closed world.

To tackle both shortcomings, we proposed CITRUN. CITRUN follows a two-step process, with mention detection and entity typing. For entity typing, we adapted the unsupervised prompting method of PromptORE for named entity recognition. We then complemented it by the use of off-domain labeled data, either coming from widely used datasets (CoNLL-2003 [39]) or synthetically annotated data (Pile-NER [40]). This off-domain data was employed to improve entity type embedding using contrastive learning. We experimentally observed that this embedding refinement step was beneficial, even when the domains were conceptually and stylistically very far away. For the mention detection subtask, we observed that the simple BIO labeling architecture was very effective, while more complex span-based models had lower cross-domain capabilities.

Experimental results showed that CITRUN significantly outperformed LLM-based unsupervised and open-world NER models while being 70 times smaller. Compared to the more supervised zero-shot NERs, CITRUN was not out of the picture and achieved competitive results. The fact that CITRUN, with its small and simple architecture, outperforms LLM-based models, particularly in low-resource settings (usually the private ground of LLMs), is of particular

[1] We recall that PromptORE was presented before the rise of LLMs (in fact, before the presentation of ChatGPT). Prompting, especially of EncLMs, was not a standard practice in the unsupervised RE domain.

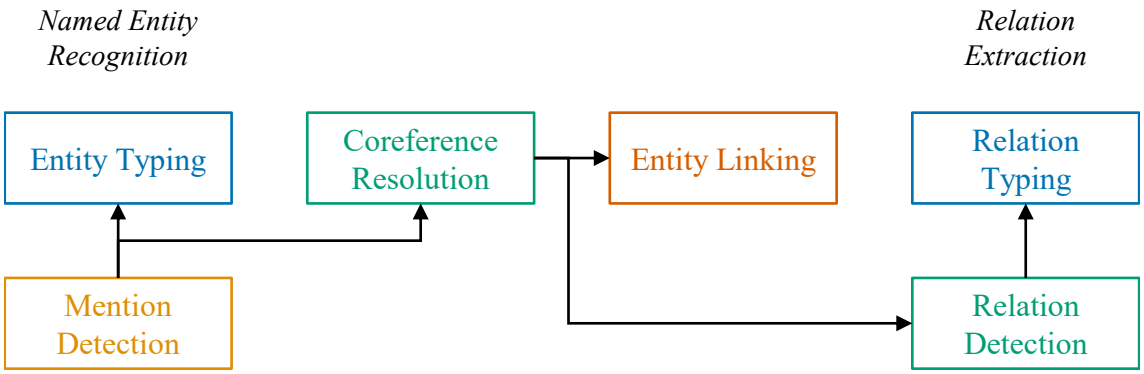


Figure VI.1: End-to-end open-world information extraction tentative architecture.

interest to the scientific community. Like PromptORE, the qualitative analysis demonstrated that CITRUN organizes the extracted entity mentions in clusters close to the true entity types.

We want to finish this summary of our contributions by returning to the motto mentioned at the beginning of this thesis: 枯れた技術の水平思考 [Lateral Thinking with Seasoned Technology]. Our three central contributions show that proven tools: the Needleman-Wunsch algorithm (1970 [132]), k-means (1957 [258, 259]), BIC (1978 [299]), ternary search, BERT (2019 [6]), BIO labeling (1999 [5]), or the triplet margin loss (2010 [300]), achieve state-of-the-art results in multiple natural language processing tasks, when they are well-chosen and carefully combined. In an era where models are exploding in size and where some mathematically and conceptually complex contributions lead to little significant results, we hope to have brought new (in fact, old) ideas to the information extraction field.

VI.2 Perspectives for Future Work

The conclusion of this research work is also the realization that although many contributions have been made, the road to efficient open-world, low-resource, and document-level IE is still long. We want to finish this thesis by succinctly evoking promising research areas for future work.

VI.2.1 Generalizing CITRUN and PromptORE to End-To-End Information Extraction

As highlighted in the literature review chapter, no open-world end-to-end IE model is available today. Implementing and testing such an end-to-end model is a priority to know its behaviors, capabilities, performances, and weaknesses to stimulate research in that area. Linked-DocRED is an ideal dataset to test and evaluate such a model, with its scale and diversity. Our two contributions with CITRUN and PromptORE for open-world RE and NER pave the way towards this objective. We propose a tentative architecture in figure VI.1.

Entity and relation typing modules can be directly adapted from PromptORE and CITRUN and have demonstrated strong performance. Mention detection is also implemented by CITRUN.

We believe a single model can tackle the coreference resolution and relation detection phases. Indeed, coreference can be seen as a particular type of relation. In fact, they both encounter the same challenges, such as extended context, combinatorial explosion, and scarcity [27, 57, 58, 59, 176, 178]. We would favor recent sequence-based (or evidence-based) document-level RE models, such as [178, 182, 184], due to their simplicity and performance. We believe they could be trained cross-domain or using synthetically annotated documents (as a side note, Li et al. [184] obtains good results with synthetically augmented datasets), extending the principles of CITRUN.

Finally, the simple entity linking baseline implemented in chapter III (Linked-DocRED) can be employed, but recent alternatives, such as ReFinED [315], are also valid choices.

VI.2.2 Combining Closed-World and Open-World Information Extraction

When analyzing a sizeable domain-specific document base, it is (nearly) impossible to determine all entity and relation types exhaustively. Usual closed-world models would miss such non-envisioned knowledge. On the contrary, with their autostructuration capabilities, open-world models can detect and organize such novel information. However, this openness comes with the drawback that the model determines the schema, and it may be more or less in line with the user's desires. Indeed, the qualitative analyses of PromptORE and CITRUN showed that they produce semantically coherent typing schemes, but they are not identical to the target ones. For instance, PromptORE merged all family relation types (see figure IV.3) or CITRUN *persons* and *scientists* (same for *artists* or *politicians*, see figure V.8). This is problematic for use cases that require a specific typing scheme to be applicable.

That is why we think a relevant research area is to combine the closed-world and open-world settings. This would allow the user to predefine a typing scheme for entities and relations he is aware of while leaving the door open to novel unseen knowledge, for which the model will provide a generated typing structure.

Some work has been done towards that topic in the open-world RE domain with RoCORE [198], CaPL [199] (soft version of RoCORE), or ASCORE [202][2]. RoCORE and ASCORE assume access to annotated data for predefined relation types. They use this data to learn a classifier to predict these relation types and improve relation embeddings for unseen relation types (that are then clustered). In practice, this supervision allows them to perform better than PromptORE on seen and unseen relations. ASCORE adopts a semi-supervised clustering approach, where labels constrain the clustering. However, these proposed approaches are not ideal as they require annotations for specific relation types on the target domain. They cannot be considered low-resource for the closed-world part. To our knowledge, no attempt has been made to combine closed-world and open-world IE in a completely low-resource setting.

A good research direction may be to generalize these previous works to be trained on cross-domain or synthetic datasets. However, a way to specify the predefined scheme during predictions would be required, as it may differ from the cross-domain train dataset.

[2] CaPL and ASCORE are anterior to PromptORE.

VI.2.3 Involve the User (in the Loop)

Finally, even when specifying the scheme beforehand, IE models will unlikely achieve perfect performances, specifically due to the low-resource constraint.

We believe another important research direction is to involve the user in the training loop so that it can correct the model's errors and refine its predictions. At the same time, we want this involvement to be minimal and the most beneficial possible. Regarding this topic, two current research areas in IE are active learning and data augmentation.

ASCORE [202] proposes an active learning strategy for RE, where the model chooses batches of relations candidates that need to be annotated by the user. The batches are selected to cover the embedding space (diversity-based active learning), and individual points are chosen to be located in high-density areas (embeddings in a small neighborhood are expected to have the same relation type). The annotated instances are then employed to constrain the clustering, which generates, in turn, pseudo labels for every relation candidate. The embedding model's weights are updated to improve the representations. However, multiple remaining questions with this approach remain open: How can we generalize it to entities, coreferences, and entity linking? How can we minimize the user's action in the annotation process? Is it better to annotate full documents, individual candidates, etc.?

Regarding the second data augmentation area, Dagdelen et al. [25] proposes a human-in-the-loop bootstrapping process for generative IE. They implement a three-step training procedure. First, the user manually annotates a small document sample (approximately 100 documents), which is used to fine-tune an LLM partially. Second, the LLM generates pseudo labels on more documents (approximately 500 in their setup) that are checked and complemented by the user. Finally, the model is wholly fine-tuned on the total sample. Their method is very promising and targets specific domains where annotation is scarce (biomedical in their paper), but it needs to be extended to open-world IE. Additionally, it could be interesting to include active learning strategies to select valuable documents to annotate more effectively.

In conclusion, involving the user in the prediction process is beginning to achieve promising results, but much work remains to be done. We believe this type of approach will be particularly interesting for the industrial domain[3] to provide tailored solutions to complex and domain-specific use cases.

[3] As a matter of fact, Dagdelen et al. [25] presents a realistic biomedical information extraction use case in their article.

Appendices

A Unsupervised Confusion Matrix

A useful tool to qualitatively analyze the performance of a classifier is the confusion matrix [316]. Each row of the confusion matrix represents the instances in an actual class (e.g., entity or relation type), and each column represents the instances in a predicted class. Thus, the matrix's diagonal shows correctly predicted instances, and the lower and upper triangles display the errors (also called confusions).

However, when implementing models based on unsupervised approaches (typically clustering), where classes are not predefined, a confusion matrix is harder to interpret. Indeed, contrary to the supervised case, there is no direct link between the class IDs and the cluster IDs (meaning the first class does not necessarily correspond to the first cluster), so there is no clear interpretable diagonal by default. To improve the readability and interoperability of a clustering confusion matrix, rows and columns must be reordered to display a diagonal and group the confusions together.

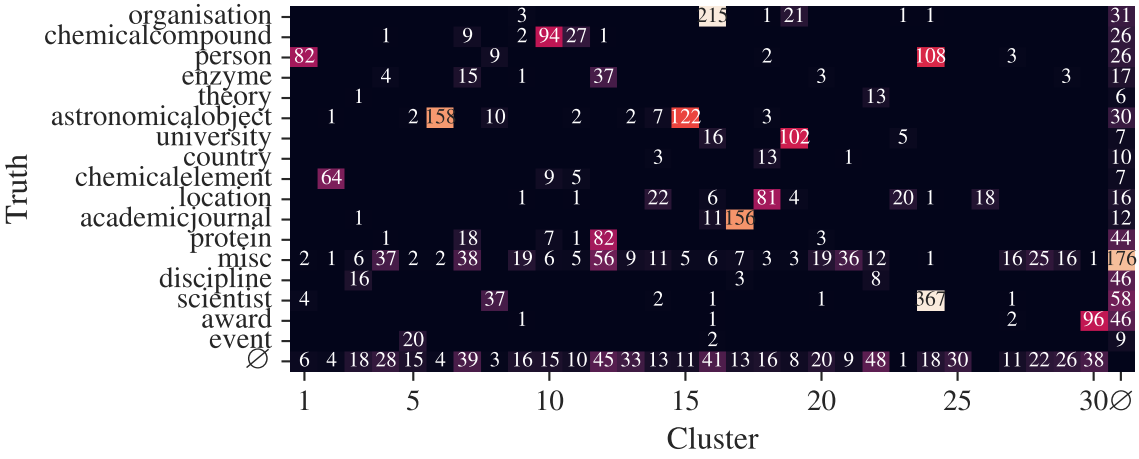
This appendix details the method employed to reorder the rows and columns. We take the example of the second figure of figure V.8 (CITRUN trained on Pile-NER and tested on Science). The initial confusion matrix, without processing, is displayed in figure A.1a. It resembles a starry sky more than a confusion matrix and is nearly impossible to interpret.

Diagonal Elicitation The first step is to find a diagonal in the confusion matrix. In a supervised scenario, if the model performs correctly, most instances are in the diagonal as the model correctly predicts them. By extension, we want to reorder the axes so that the unsupervised confusion matrix shows a clear diagonal: we want to find the "main" cluster corresponding to each class. For instance, in figure A.1a, most instances of *organization* are in cluster 16, most *chemicalcompound* entities are in cluster 11, ...

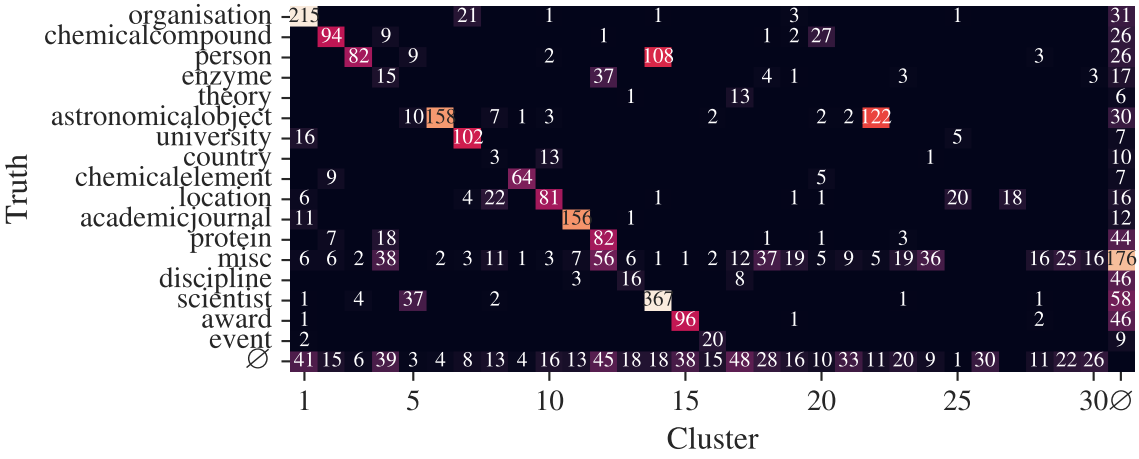
This can be formulated as: "reorganizing the rows and columns so that the diagonal of the matrix is of maximal sum". This corresponds to an assignment problem (except that the canonical problem involves minimizing the sum). We solve this assignment problem using a modified version of the Jonker-Volgenant algorithm[1] [317, 318].

The resulting confusion matrix is displayed in figure A.1b. It displays a clear diagonal that is much more interpretable than the initial confusion matrix. Nevertheless, some important values outside the diagonal are still scattered (e.g., *person*/cluster 14, *astronomicalobject*/cluster 22).

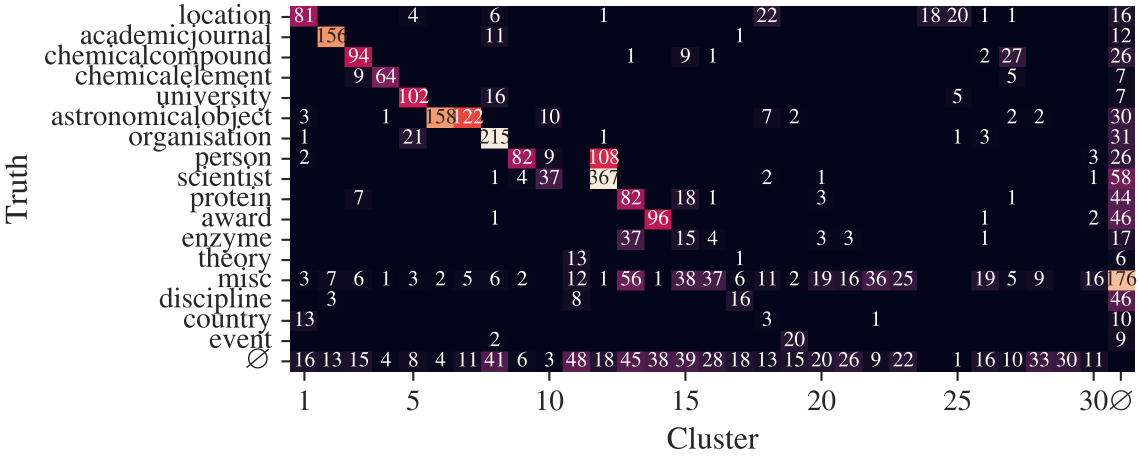
[1]We employ the SciPy implementation available at https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.linear_sum_assignment.html (visited on 2024-10-21).



(a) Raw confusion matrix.



(b) After diagonal elicitation.



(c) After confusion grouping.

Figure A.1: Reordering of the unsupervised confusion matrix of CITRUN for NER trained on Pile-NER and tested on Science.

Confusion Grouping The second step aims to bring major confusions closer to make the matrix readable. An ideal confusion matrix is a band matrix, that is, a sparse matrix where the non-zero entries are confined to a diagonal band. We propose implementing the reverse Cutthill–McKee algorithm[2] [319, 320], which aims to permute a sparse matrix into a band matrix with a small bandwidth. In practice, not all non-zero values are interesting (some represent noise or very rare edge cases), so we propose fixing a threshold (1 % of the total instances). Below this threshold, the value is not considered when reordering axes.

We obtain the final confusion matrix of figure A.1c. We can see that the major confusions are now grouped closer (e.g., *person* and *scientist*, *protein* and *enzyme*, *university* and *organization*).

As a side note, the first diagonal elicitation step is optional, as the reverse Cuthill–McKee algorithm produces a band matrix (that is, with a diagonal). We have found, in practice, that the first diagonal elicitation step helped to produce a diagonal with the maximum sum, thus leading to a clearer interpretation.

[2]Following the SciPy implementation available at https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.csgraph.reverse_cuthill_mckee.html (visited on 2024-10-21).

Bibliography

[1] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. "DocRED: A large-scale document-level relation extraction dataset". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019, pp. 764–777. ISBN: 978-1-950737-48-2. DOI: 10.18653/v1/p19-1074 (cit. on pp. ii, iii, 4, 6, 12, 19, 20, 26, 27, 28, 29, 30, 35, 41, 43, 52).

[2] Claude Elwood Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (July 1948), pp. 379–423. ISSN: 0005-8580. DOI: 10.1002/j.1538-7305.1948.tb01338.x. URL: https://ieeexplore.ieee.org/abstract/document/6773024 (cit. on pp. xii, 49, 50).

[3] Lawrence Hubert and Phipps Arabie. "Comparing partitions". In: *Journal of Classification* 2.1 (Dec. 1985). Publisher: Springer, pp. 193–218. DOI: 10.1007/BF01908075. URL: https://link.springer.com/article/10.1007/BF01908075 (cit. on pp. xii, 49, 58).

[4] Douglas Steinley. "Properties of the Hubert-Arabie adjusted Rand index". In: *Psychological Methods* 9.3 (Sept. 2004). Publisher: Psychol Methods, pp. 386–396. DOI: 10.1037/1082-989X.9.3.386. URL: https://pubmed.ncbi.nlm.nih.gov/15355155/ (cit. on pp. xii, 49, 58).

[5] Lance Ramshaw and Mitchell Marcus. "Text Chunking Using Transformation-Based Learning". In: *Natural Language Processing Using Very Large Corpora*. Ed. by Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky. Text, Speech and Language Technology. Dordrecht: Springer Netherlands, 1999, pp. 157–176. ISBN: 978-94-017-2390-9. DOI: 10.1007/978-94-017-2390-9_10. URL: https://doi.org/10.1007/978-94-017-2390-9_10 (cit. on pp. xii, 111).

[6] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. ISBN: 978-1-950737-13-0. DOI: 10.18653/V1/N19-1423 (cit. on pp. xii, 15, 51, 57, 61, 77, 83, 90, 104, 111).

[7] Thomas N. Kipf and Max Welling. *Semi-Supervised Classification with Graph Convolutional Networks*. Feb. 22, 2017. DOI: 10.48550/arXiv.1609.02907. arXiv: 1609.02907[cs, stat]. URL: http://arxiv.org/abs/1609.02907 (cit. on pp. xii, 15, 20).

[8] A. Sperduti and A. Starita. "Supervised neural networks for the classification of structures". In: *IEEE Transactions on Neural Networks* 8.3 (May 1997). Conference Name: IEEE Transactions on Neural Networks, pp. 714–735. ISSN: 1941-0093. DOI: 10.1109/72.572108. URL: https://ieeexplore.ieee.org/abstract/document/572108 (cit. on pp. xii, 14, 15).

[9] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (Nov. 1997). Conference Name: Neural Computation, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://ieeexplore.ieee.org/abstract/document/6795963 (cit. on pp. xii, 14).

[10] Central Secretary ISO. *Quantities and units — Part 2: Mathematical signs and symbols to be used in the natural sciences and technology*. Version 2019. Geneva, CH, 2019. URL: https://www.iso.org/standard/64973.html (cit. on p. xiv).

[11] Taku Kudo and John Richardson. "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Eduardo Blanco and Wei Lu. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. DOI: 10.18653/v1/D18-2012. URL: https://aclanthology.org/D18-2012 (cit. on pp. xiv, 11).

[12] Petroc Taylor. *Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025*. 871513. Statista, 2023. URL: https://www.statista.com/statistics/871513/worldwide-data-created/ (cit. on p. 1).

[13] Shubhangi Vashisth, Erick Brethenoux, Stephen Emmott, Melissa Davis, David Norrie, and Bern Elliot. *Market Guide for Text Analytics*. Gartner research, 2020. URL: https://www.gartner.com/en/documents/3989657 (cit. on p. 1).

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space*. Sept. 6, 2013. DOI: 10.48550/arXiv.1301.3781. arXiv: 1301.3781[cs]. URL: http://arxiv.org/abs/1301.3781 (cit. on pp. 2, 14).

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Proceedings of Advances in Neural Information Processing Systems 30*. 31st Conference on Neural Information Processing Systems. Vol. 2017-Decem. Long Beach, CA, USA: Neural information processing systems foundation, June 2017, pp. 5999–6009. DOI: 10.48550/arxiv.1706.03762. URL: https://arxiv.org/abs/1706.03762v5 (cit. on pp. 2, 15, 57, 61).

[16] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. *LLaMA: Open and Efficient Foundation Language Models*. Feb. 27, 2023. DOI: 10.48550/arXiv.2302.13971. arXiv: 2302.13971[cs]. URL: http://arxiv.org/abs/2302.13971 (cit. on pp. 2, 17, 80).

[17] Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. "A Comprehensive Survey on Automatic Knowledge Graph Construction". In: *ACM Comput. Surv.* 56.4 (Nov. 30, 2023), 94:1–94:62. ISSN: 0360-0300. DOI: 10.1145/3618295. URL: https://doi.org/10.1145/3618295 (cit. on p. 3).

[18] Hanwen Zheng, Sijia Wang, and Lifu Huang. *A Survey of Document-Level Information Extraction*. Sept. 23, 2023. DOI: 10.48550/arXiv.2309.13249. arXiv: 2309.13249[cs]. URL: http://arxiv.org/abs/2309.13249 (cit. on pp. 3, 10).

[19] Shaowen Zhou, Bowen Yu, Aixin Sun, Cheng Long, Jingyang Li, and Jian Sun. "A Survey on Neural Open Information Extraction: Current Status and Future Directions". In: Thirty-First International Joint Conference on Artificial Intelligence. Vol. 6. ISSN: 1045-0823. July 16, 2022, pp. 5694–5701. DOI: 10.24963/ijcai.2022/793. URL: https://www.ijcai.org/proceedings/2022/793 (cit. on pp. 3, 10).

[20] Severine Verlinden, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. "Injecting Knowledge Base Information into End-to-End Joint Entity and Relation Extraction and Coreference Resolution". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Findings 2021. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, Aug. 2021, pp. 1952–1957. DOI: 10.18653/v1/2021.findings-acl.171. URL: https://aclanthology.org/2021.findings-acl.171 (cit. on pp. 3, 10, 11, 14, 21, 28, 51, 52).

[21] Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. "DWIE: An entity-centric dataset for multi-task document-level information extraction". In: *Information Processing & Management* 58.4 (July 2021), p. 102563. ISSN: 03064573. DOI: 10.1016/j.ipm.2021.102563. URL: https://linkinghub.elsevier.com/retrieve/pii/S0306457321000662 (cit. on pp. 3, 6, 25, 26, 27, 28, 29, 30, 46, 47, 51, 52, 109).

[22] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. "Entity, relation, and event extraction with contextualized span representations". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 5784–5789. ISBN: 978-1-950737-90-1. DOI: 10.18653/v1/d19-1585 (cit. on pp. 3, 11, 15, 28, 57, 61).

[23] Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. "GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction". In: *Proceedings of the Twelfth International Conference on Learning Representations*. The Twelfth International Conference on Learning Representations. Vienna, Austria, Jan. 24, 2024. URL: https://openreview.net/forum?id=Y3wpuxd7u9 (cit. on pp. 3, 10, 17, 18, 19, 78, 80, 81, 87, 90).

[24] Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. "Unified Structure Generation for Universal Information Extraction". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2022. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5755–5772. DOI: 10.18653/v1/2022.acl-long.395. URL: https://aclanthology.org/2022.acl-long.395 (cit. on pp. 3, 17, 18).

[25] John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. "Structured information extraction from scientific text with large language models". In: *Nature Communications* 15.1 (Feb. 15, 2024). Publisher: Nature Publishing Group, p. 1418. ISSN: 2041-1723. DOI: 10.1038/s41467-024-45563-x. URL: https://www.nature.com/articles/s41467-024-45563-x (cit. on pp. 3, 19, 113).

[26] Benjamin Townsend, Eamon Ito-Fisher, Lily Zhang, and Madison May. *Doc2Dict: Information Extraction as Text Generation*. Oct. 10, 2021. DOI: 10.48550/arXiv.2105.07510. arXiv: 2105.07510[cs]. URL: http://arxiv.org/abs/2105.07510 (cit. on pp. 3, 17).

[27] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. "End-to-end Neural Coreference Resolution". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2017. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 188–197. DOI: 10.18653/v1/D17-1018. URL: https://aclanthology.org/D17-1018 (cit. on pp. 5, 12, 14, 112).

[28] Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. "Self-Improving for Zero-Shot Named Entity Recognition with Large Language Models". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. NAACL-HLT 2024. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 583–593. URL: https://aclanthology.org/2024.naacl-short.49 (cit. on pp. 5, 16, 17, 18, 88).

[29] Ethan Perez, Douwe Kiela, and Kyunghyun Cho. "True Few-Shot Learning with Language Models". In: (May 2021) (cit. on pp. 5, 55, 57, 58).

[30] OpenAI. *ChatGPT — Release Notes | OpenAI Help Center*. 2022. URL: https://help.openai.com/en/articles/6825453-chatgpt-release-notes (cit. on pp. 6, 59).

[31] Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. "UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition". In: *Proceedings of the Twelfth International Conference on Learning Representations*. The Twelfth International Conference on Learning Representations. Vienna, Austria, 2024. URL: https://openreview.net/forum?id=r65xfUb76p (cit. on pp. 6, 16, 17, 18, 77, 78, 80, 81, 83, 87, 88, 89, 90).

[32] Pierre-Yves Genest, Pierre-Edouard Portier, Előd Egyed-Zsigmond, and Martino Lovisetto. “Linked-DocRED – Enhancing DocRED with Entity-Linking to Evaluate End-To-End Document-Level Information Extraction Pipelines”. In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR’23. Taipei, Taiwan: Association for Computing Machinery, 2023. ISBN: 978-1-4503-9408-6. DOI: 10.1145/3539618.3591912 (cit. on pp. 6, 25).

[33] Pierre-Yves Genest, Pierre-Edouard Portier, Előd Egyed-Zsigmond, and Laurent-Walter Goix. “PromptORE - A Novel Approach Towards Fully Unsupervised Relation Extraction”. In: *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*. CIKM ’22. Atlanta, USA: Association for Computing Machinery, Oct. 17, 2022. DOI: 10.1145/3511808.3557422. URL: https://hal.science/hal-03858264 (cit. on pp. 7, 54, 84).

[34] Pierre-Yves Genest, Pierre-Edouard Portier, Előd Egyed-Zsigmond, and Laurent-Walter Goix. “PromptORE – Vers l’Extraction de Relations non-supervisée”. In: *Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles*. 30e Conférence sur le Traitement Automatique des Langues Naturelles. Vol. 4. Paris, France, June 5, 2023, pp. 58–64. URL: https://coria-taln-2023.sciencesconf.org/459305 (cit. on p. 7).

[35] Chen Jia, Xiaobo Liang, and Yue Zhang. “Cross-Domain NER using Cross-Domain Language Modeling”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2464–2474. DOI: 10.18653/v1/P19-1236. URL: https://aclanthology.org/P19-1236 (cit. on pp. 8, 81, 110).

[36] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. “A Unified MRC Framework for Named Entity Recognition”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 5849–5859. DOI: 10.18653/v1/2020.acl-main.519. URL: https://aclanthology.org/2020.acl-main.519 (cit. on pp. 8, 15, 110).

[37] Qi Peng, Changmeng Zheng, Yi Cai, Tao Wang, Haoran Xie, and Qing Li. “Unsupervised cross-domain named entity recognition using entity-aware adversarial training”. In: *Neural Networks* 138 (June 1, 2021), pp. 68–77. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2020.12.027. URL: https://www.sciencedirect.com/science/article/pii/S0893608020304524 (cit. on pp. 8, 77, 81, 110).

[38] Andrea Iovine, Anjie Fang, Besnik Fetahu, Oleg Rokhlenko, and Shervin Malmasi. “CycleNER: An Unsupervised Training Approach for Named Entity Recognition”. In: *Proceedings of the ACM Web Conference 2022*. WWW ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 2916–2924. ISBN: 978-1-4503-9096-5. DOI: 10.1145/3485447.3512012. URL: https://dl.acm.org/doi/10.1145/3485447.3512012 (cit. on pp. 8, 81, 110).

[39] Erik Tjong Kim Sang and Fien De Meulder. “Introduction to the CoNLL-2003 shared task: language-independent named entity recognition”. In: *Proceedings of the 7th conference on Natural language learning*. HLT-NAACL 2003. Vol. Volume 4. CONLL ’03. USA: Association for Computational Linguistics, 2003, pp. 142–147. DOI: 10.3115/1119176.1119195. URL: https://dl.acm.org/doi/10.3115/1119176.1119195 (cit. on pp. 8, 12, 77, 78, 89, 110).

[40] Jie Lou, Yaojie Lu, Dai Dai, Wei Jia, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. “Universal Information Extraction as Unified Semantic Matching”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.11 (June 26, 2023). Number: 11, pp. 13318–13326. ISSN: 2374-3468. DOI: 10.1609/aaai.v37i11.26563. URL: https://ojs.aaai.org/index.php/AAAI/article/view/26563 (cit. on pp. 8, 15, 16, 17, 110).

[41] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. “A Survey on Knowledge Graph-Based Recommender Systems”. In: *IEEE Transactions on Knowledge and Data Engineering* 34.8 (Aug. 1, 2022). Publisher: Institute of Electrical and Electronics Engineers (IEEE), pp. 3549–3568. DOI: 10.1109/tkde.2020.3028705 (cit. on p. 10).

[42] Xiaojun Chen, Shengbin Jia, and Yang Xiang. “A review: Knowledge reasoning over knowledge graph”. In: *Expert Systems with Applications* 141 (Mar. 2020). Publisher: Pergamon, pp. 112948–112948. DOI: 10.1016/j.eswa.2019.112948 (cit. on p. 10).

[43] Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. “Knowledge graph embedding based question answering”. In: *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*. Melbourne, Australia: Association for Computing Machinery, Inc, Jan. 2019, pp. 105–113. ISBN: 978-1-4503-5940-5. DOI: 10.1145/3289600.3290956 (cit. on p. 10).

[44] Maxime Prieur, Cédric Du Mouza, Guillaume Gadek, and Bruno Grilhères. “Peuplement de base de connaissances, liage dynamique et système end-to-end”. In: *Revue des Nouvelles Technologies de l’Information* Extraction et Gestion des Connaissances, RNTI-E-39 (2023), pp. 281–288. URL: https://hal.science/hal-03887658 (cit. on pp. 10, 11, 28, 51, 52).

[45] Sarah Elhammadi, Laks V.S. Lakshmanan, Raymond Ng, Michael Simpson, Baoxing Huai, Zhefeng Wang, and Lanjun Wang. “A High Precision Pipeline for Financial Knowledge Graph Construction”. In: *Proceedings of the 28th International Conference on Computational Linguistics*. COLING 2020. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 967–977. DOI: 10.18653/v1/2020.coling-main.84. URL: https://aclanthology.org/2020.coling-main.84 (cit. on pp. 10, 28).

[46] Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. “CodeIE: Large Code Generation Models are Better Few-Shot Information Extractors”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2023. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for

Computational Linguistics, July 2023, pp. 15339–15353. DOI: 10.18653/v1/2023.acl-long.855. URL: https://aclanthology.org/2023.acl-long.855 (cit. on pp. 10, 11, 17).

[47] Keming Lu, Xiaoman Pan, Kaiqiang Song, Hongming Zhang, Dong Yu, and Jianshu Chen. "PIVOINE: Instruction Tuning for Open-world Entity Profiling". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Findings 2023. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 15108–15127. DOI: 10.18653/v1/2023.findings-emnlp.1009. URL: https://aclanthology.org/2023.findings-emnlp.1009 (cit. on pp. 10, 17, 22).

[48] Zixuan Li, Yutao Zeng, Yuxin Zuo, Weicheng Ren, Wenxuan Liu, Miao Su, Yucan Guo, Yantao Liu, Xiang Li, Zhilei Hu, Long Bai, Wei Li, Yidan Liu, Pan Yang, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. *KnowCoder: Coding Structured Knowledge into LLMs for Universal Information Extraction*. Mar. 13, 2024. DOI: 10.48550/arXiv.2403.07969. arXiv: 2403.07969[cs]. URL: http://arxiv.org/abs/2403.07969 (cit. on pp. 10, 18).

[49] Ryan Clancy, Ihab F. Ilyas, and Jimmy Lin. "Scalable Knowledge Graph Construction from Text Collections". In: *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*. Ed. by James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 39–46. DOI: 10.18653/v1/D19-6607. URL: https://aclanthology.org/D19-6607 (cit. on pp. 10, 14).

[50] Xinglin Xiao, Yijie Wang, Nan Xu, Yuqi Wang, Hanxuan Yang, Minzheng Wang, Yin Luo, Lei Wang, Wenji Mao, and Daniel Zeng. *YAYI-UIE: A Chat-Enhanced Instruction Tuning Framework for Universal Information Extraction*. Apr. 2, 2024. DOI: 10.48550/arXiv.2312.15548. arXiv: 2312.15548[cs]. URL: http://arxiv.org/abs/2312.15548 (cit. on pp. 10, 17, 18).

[51] Pai Liu, Wenyang Gao, Wenjie Dong, Lin Ai, Ziwei Gong, Songfang Huang, Zongsheng Li, Ehsan Hoque, Julia Hirschberg, and Yue Zhang. *A Survey on Open Information Extraction from Rule-based Model to Large Language Model*. May 10, 2024. DOI: 10.48550/arXiv.2208.08690. arXiv: 2208.08690[cs]. URL: http://arxiv.org/abs/2208.08690 (cit. on pp. 10, 66).

[52] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. *Large Language Models for Generative Information Extraction: A Survey*. June 4, 2024. DOI: 10.48550/arXiv.2312.17617. arXiv: 2312.17617[cs]. URL: http://arxiv.org/abs/2312.17617 (cit. on pp. 10, 21).

[53] Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. "Web-scale information extraction in knowitall: (preliminary results)". In: *Proceedings of the 13th international conference on World Wide Web*. WWW '04. New York, NY, USA: Association for Computing Machinery, 2004, pp. 100–110. ISBN: 978-1-58113-844-3. DOI: 10.1145/988672.988687. URL: https://doi.org/10.1145/988672.988687 (cit. on pp. 11, 13, 22).

[54] Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. “Open Information Extraction from the Web”. In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. Hyderabad, India: Morgan Kaufmann Publishers Inc., 2007, pp. 2670–2676 (cit. on pp. 11, 13, 22).

[55] Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. “Supervised open information extraction”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1. New Orleans, Louisiana, United States: Association for Computational Linguistics, 2018, pp. 885–895. ISBN: 978-1-948087-27-8. DOI: 10.18653/v1/n18-1081 (cit. on pp. 11, 14, 22).

[56] Lei Cui, Furu Wei, and Ming Zhou. “Neural open information extraction”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Vol. 2. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 407–413. ISBN: 978-1-948087-34-6. DOI: 10.18653/v1/p18-2065. URL: https://aclanthology.org/P18-2065 (cit. on pp. 11, 14, 16, 21, 22).

[57] Yuval Kirstain, Ori Ram, and Omer Levy. “Coreference Resolution without Span Representations”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. ACL-IJCNLP 2021. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, Aug. 2021, pp. 14–19. DOI: 10.18653/v1/2021.acl-short.3. URL: https://aclanthology.org/2021.acl-short.3 (cit. on pp. 11, 12, 112).

[58] Vladimir Dobrovolskii. “Word-Level Coreference Resolution”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2021. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7670–7675. DOI: 10.18653/v1/2021.emnlp-main.605. URL: https://aclanthology.org/2021.emnlp-main.605 (cit. on pp. 11, 12, 79, 97, 112).

[59] Kenton Lee, Luheng He, and Luke Zettlemoyer. “Higher-Order Coreference Resolution with Coarse-to-Fine Inference”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. NAACL-HLT 2018. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 687–692. DOI: 10.18653/v1/N18-2108. URL: https://aclanthology.org/N18-2108 (cit. on pp. 11, 12, 14, 112).

[60] Yilun Zhu, Siyao Peng, Sameer Pradhan, and Amir Zeldes. “Incorporating Singletons and Mention-based Features in Coreference Resolution via Multi-task Learning for Better Generalization”. In: *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*. IJCNLP-AACL 2023. Ed. by Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi. Nusa Dua, Bali: Association for Computational

Linguistics, Nov. 2023, pp. 121–130. URL: https://aclanthology.org/2023.ijcnlp-short.14 (cit. on p. 11).

[61] Klim Zaporojets, Johannes Deleu, Yiwei Jiang, Thomas Demeester, and Chris Develder. "Towards Consistent Document-level Entity Linking: Joint Models for Entity Linking and Coreference Resolution". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL 2022. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 778–784. DOI: 10.18653/v1/2022.acl-short.88. URL: https://aclanthology.org/2022.acl-short.88 (cit. on pp. 11, 13).

[62] Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. "Entity Linking and Discovery via Arborescence-based Supervised Clustering". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2022, pp. 4644–4658. DOI: 10.18653/v1/2022.naacl-main.343. arXiv: 2109.01242[cs]. URL: http://arxiv.org/abs/2109.01242 (cit. on pp. 11, 13).

[63] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. "An Autoregressive Text-to-Graph Framework for Joint Entity and Relation Extraction". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.17 (Mar. 24, 2024). Number: 17, pp. 19477–19487. ISSN: 2374-3468. DOI: 10.1609/aaai.v38i17.29919. URL: https://ojs.aaai.org/index.php/AAAI/article/view/29919 (cit. on pp. 11, 16, 17, 19).

[64] Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. "A general framework for information extraction using dynamic span graphs". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 3036–3046. ISBN: 978-1-950737-13-0. DOI: 10.18653/v1/n19-1308 (cit. on pp. 11, 14, 15, 19, 56, 57).

[65] Urchade Zaratiana, Nadi Tomeh, Yann Dauxais, Pierre Holat, and Thierry Charnois. *EnriCo: Enriched Representation and Globally Constrained Inference for Entity and Relation Extraction*. Apr. 18, 2024. DOI: 10.48550/arXiv.2404.12493. arXiv: 2404.12493[cs]. URL: http://arxiv.org/abs/2404.12493 (cit. on pp. 11, 15).

[66] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. *GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer*. Nov. 14, 2023. DOI: 10.48550/arXiv.2311.08526. arXiv: 2311.08526[cs]. URL: http://arxiv.org/abs/2311.08526 (cit. on pp. 12, 16, 19, 45, 78, 80, 81, 83, 87, 88).

[67] Amir Zeldes and Shuo Zhang. "When Annotation Schemes Change Rules Help: A Configurable Approach to Coreference Resolution beyond OntoNotes". In: *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*. CORBON 2016. Ed. by Maciej Ogrodniczuk and Vincent Ng. San Diego, California: Association for Computational Linguistics, June 2016, pp. 92–101. DOI: 10.18653/v1/W16-0713. URL: https://aclanthology.org/W16-0713 (cit. on p. 12).

[68] Ian Porada, Alexandra Olteanu, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. *Investigating Failures to Generalize for Coreference Resolution Models*. Mar. 16, 2023. DOI: 10.48550/arXiv.2303.09092. arXiv: 2303.09092[cs]. URL: http://arxiv.org/abs/2303.09092 (cit. on p. 12).

[69] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. "Towards Robust Linguistic Analysis using OntoNotes". In: *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. CoNLL 2013. Ed. by Julia Hockenmaier and Sebastian Riedel. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 143–152. URL: https://aclanthology.org/W13-3516 (cit. on p. 12).

[70] Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. "PreCo: A Large-scale Dataset in Preschool Vocabulary for Coreference Resolution". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2018. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 172–181. DOI: 10.18653/v1/D18-1016. URL: https://aclanthology.org/D18-1016 (cit. on p. 12).

[71] Wei Shen, Jianyong Wang, and Jiawei Han. "Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions". In: *IEEE Transactions on Knowledge and Data Engineering* 27.2 (Feb. 1, 2015), pp. 443–460. ISSN: 1041-4347. DOI: 10.1109/TKDE.2014.2327028. URL: http://ieeexplore.ieee.org/document/6823700/ (cit. on p. 13).

[72] Heng Ji and Ralph Grishman. "Knowledge base population: successful approaches and challenges". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. HLT '11. USA: Association for Computational Linguistics, 2011, pp. 1148–1158. ISBN: 978-1-932432-87-9 (cit. on p. 13).

[73] Tom Ayoola, Joseph Fisher, and Andrea Pierleoni. "Improving Entity Disambiguation by Reasoning over a Knowledge Base". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2022. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 2899–2912. DOI: 10.18653/v1/2022.naacl-main.210. URL: https://aclanthology.org/2022.naacl-main.210 (cit. on p. 13).

[74] Xianpei Han, Le Sun, and Jun Zhao. "Collective entity linking in web text: a graph-based method". In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. SIGIR '11. New York, NY, USA: Association for Computing Machinery, 2011, pp. 765–774. ISBN: 978-1-4503-0757-4. DOI: 10.1145/2009916.2010019. URL: https://doi.org/10.1145/2009916.2010019 (cit. on p. 13).

[75] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. "Translating Embeddings for Modeling Multi-relational Data". In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc., 2013. URL:

https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html (cit. on pp. 13, 14).

[76] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. "Cross-Sentence N-ary Relation Extraction with Graph LSTMs". In: *Transactions of the Association for Computational Linguistics* 5 (Apr. 1, 2017), pp. 101–115. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00049. URL: https://doi.org/10.1162/tacl_a_00049 (cit. on pp. 13, 14, 20).

[77] Beth M Sundheim. "Overview of the third message understanding evaluation and conference". In: *Third message understanding conference (MUC-3): Proceedings of a conference held in san diego, california, may 21-23, 1991*. 1991 (cit. on p. 13).

[78] Beth M Sundheim. "Overview of the fourth message understanding evaluation and conference". In: *Fourth message understanding conference (MUC-4): Proceedings of a conference held in McLean, virginia, june 16-18, 1992*. 1992 (cit. on p. 13).

[79] Beth M Sundheim. "Overview of results of the MUC-6 evaluation". In: *Sixth message understanding conference (MUC-6): Proceedings of a conference held in columbia, maryland, november 6-8, 1995*. 1995 (cit. on p. 13).

[80] Nancy Chinchor and Patricia Robinson. "MUC-7 named entity task definition". In: *Proceedings of the 7th conference on message understanding*. Vol. 29. 1997, pp. 1–21 (cit. on p. 13).

[81] Nancy Chinchor. "Overview of MUC-7". In: *Seventh message understanding conference (MUC-7): Proceedings of a conference held in fairfax, virginia, april 29-may 1, 1998*. 1998 (cit. on p. 13).

[82] Lynette Hirschman. "The evolution of evaluation: Lessons from the message understanding conferences". In: *Computer Speech & Language* 12.4 (1998), pp. 281–305 (cit. on p. 13).

[83] Gabor Angeli, Melvin Johnson Premkumar, and Christopher D. Manning. "Leveraging linguistic structure for open domain information extraction". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*. Vol. 1. Association for Computational Linguistics, 2015, pp. 344–354. ISBN: 978-1-941643-72-3. DOI: 10.3115/v1/p15-1034. URL: https://aclanthology.org/P15-1034 (cit. on pp. 13, 14).

[84] Yusuke Shinyama and Satoshi Sekine. "Preemptive Information Extraction using Unrestricted Relation Discovery". In: *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. NAACL-HLT 2006. Ed. by Robert C. Moore, Jeff Bilmes, Jennifer Chu-Carroll, and Mark Sanderson. New York City, USA: Association for Computational Linguistics, June 2006, pp. 304–311. URL: https://aclanthology.org/N06-1039 (cit. on pp. 13, 22).

[85] Fei Wu and Daniel S. Weld. "Open Information Extraction Using Wikipedia". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL 2010. Ed. by Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 118–127. URL: https://aclanthology.org/P10-1013 (cit. on pp. 14, 22).

[86] Anthony Fader, Stephen Soderland, and Oren Etzioni. "Identifying Relations for Open Information Extraction". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2011. Ed. by Regina Barzilay and Mark Johnson. Edinburgh, Scotland, UK.: Association for Computational Linguistics, July 2011, pp. 1535–1545. URL: https://aclanthology.org/D11-1142 (cit. on pp. 14, 22).

[87] Luciano Del Corro and Rainer Gemulla. "ClausIE: clause-based open information extraction". In: *Proceedings of the 22nd international conference on World Wide Web*. WWW '13. New York, NY, USA: Association for Computing Machinery, 2013, pp. 355–366. ISBN: 978-1-4503-2035-1. DOI: 10.1145/2488388.2488420. URL: https://doi.org/10.1145/2488388.2488420 (cit. on pp. 14, 22).

[88] Swarnadeep Saha and Mausam. "Open information extraction from conjunctive sentences". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, pp. 2288–2299. ISBN: 978-1-948087-50-6 (cit. on pp. 14, 22).

[89] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global vectors for word representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. ISBN: 978-1-937284-96-1. DOI: 10.3115/v1/d14-1162. URL: https://aclanthology.org/D14-1162 (cit. on pp. 14, 51, 57, 58).

[90] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5 (2017). Ed. by Lillian Lee, Mark Johnson, and Kristina Toutanova. Place: Cambridge, MA Publisher: MIT Press, pp. 135–146. DOI: 10.1162/tacl_a_00051. URL: https://aclanthology.org/Q17-1010 (cit. on p. 14).

[91] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. EMNLP 2014. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. DOI: 10.3115/v1/D14-1179. URL: https://aclanthology.org/D14-1179 (cit. on pp. 14, 16).

[92] Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. "Neural Temporal Relation Extraction". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. EACL 2017. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller.

Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 746–751. URL: https://aclanthology.org/E17-2118 (cit. on pp. 14, 15).

[93] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. "Relation Classification via Convolutional Deep Neural Network". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. COLING 2014. Ed. by Junichi Tsujii and Jan Hajic. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 2335–2344. URL: https://aclanthology.org/C14-1220 (cit. on p. 14).

[94] ChunYang Liu, WenBo Sun, WenHan Chao, and WanXiang Che. "Convolution Neural Network for Relation Extraction". In: *Advanced Data Mining and Applications*. Ed. by Hiroshi Motoda, Zhaohui Wu, Longbing Cao, Osmar Zaiane, Min Yao, and Wei Wang. Berlin, Heidelberg: Springer, 2013, pp. 231–242. ISBN: 978-3-642-53917-6. DOI: 10.1007/978-3-642-53917-6_21 (cit. on p. 14).

[95] Huiwei Zhou, Shixian Ning, Yunlong Yang, Zhuang Liu, Chengkun Lang, and Yingyu Lin. "Chemical-induced disease relation extraction with dependency information and prior knowledge". In: *Journal of Biomedical Informatics* 84 (Aug. 1, 2018), pp. 171–178. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2018.07.007. URL: https://www.sciencedirect.com/science/article/pii/S1532046418301333 (cit. on p. 14).

[96] Pengfei Li and Kezhi Mao. "Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts". In: *Expert Systems with Applications* 115 (Jan. 1, 2019), pp. 512–523. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2018.08.009. URL: https://www.sciencedirect.com/science/article/pii/S0957417418305177 (cit. on p. 14).

[97] Zhi Li, Jinshan Yang, Xu Gou, and Xiaorong Qi. "Recurrent neural networks with segment attention and entity description for relation extraction from clinical texts". In: *Artificial Intelligence in Medicine* 97 (June 1, 2019), pp. 9–18. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2019.04.003. URL: https://www.sciencedirect.com/science/article/pii/S093336571830753X (cit. on p. 14).

[98] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Ed. by Kalina Bontcheva and Jingbo Zhu. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 55–60. DOI: 10.3115/v1/P14-5010. URL: https://aclanthology.org/P14-5010 (cit. on p. 14).

[99] Yue Yuan, Xiaofei Zhou, Shirui Pan, Qiannan Zhu, Zeliang Song, and Li Guo. "A relation-specific attention network for joint entity and relation extraction". In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. IJCAI'20. Yokohama, Yokohama, Japan, Jan. 7, 2021, pp. 4054–4060. ISBN: 978-0-9992411-6-5 (cit. on p. 14).

[100] Junlang Zhan and Hai Zhao. "Span Model for Open Information Extraction on Accurate Corpus". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.5 (Apr. 3, 2020). Number: 05, pp. 9523–9530. ISSN: 2374-3468. DOI: 10.1609/aaai.v34i05.6497. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6497 (cit. on p. 14).

[101] Chris Quirk and Hoifung Poon. "Distant Supervision for Relation Extraction beyond the Sentence Boundary". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. EACL 2017. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1171–1182. URL: https://aclanthology.org/E17-1110 (cit. on pp. 14, 20).

[102] Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. "Connecting the dots: Document-level neural relation extraction with edge-oriented graphs". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*. 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 4925–4936. ISBN: 978-1-950737-90-1. DOI: 10.18653/v1/d19-1498. URL: https://aclanthology.org/D19-1498 (cit. on pp. 14, 20, 51).

[103] Bowen Yu, Zhenyu Zhang, Xiaobo Shu, Tingwen Liu, Yubin Wang, Bin Wang, and Sujian Li. "Joint Extraction of Entities and Relations Based on a Novel Decomposition Strategy". In: *ECAI 2020*. IOS Press, 2020, pp. 2282–2289. DOI: 10.3233/FAIA200356. URL: https://ebooks.iospress.nl/doi/10.3233/FAIA200356 (cit. on p. 14).

[104] Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. "A Joint Neural Model for Information Extraction with Global Features". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7999–8009. DOI: 10.18653/v1/2020.acl-main.713 (cit. on pp. 15, 56).

[105] Minh Van Nguyen, Viet Dac Lai, and Thien Huu Nguyen. "Cross-Task Instance Representation Interactions and Label Dependencies for Joint Information Extraction with Graph Convolutional Networks". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2021. Ed. by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou. Online: Association for Computational Linguistics, June 2021, pp. 27–38. DOI: 10.18653/v1/2021.naacl-main.3. URL: https://aclanthology.org/2021.naacl-main.3 (cit. on p. 15).

[106] Urchade Zaratiana, Nadi Tomeh, Niama El Khbir, Pierre Holat, and Thierry Charnois. *GraphER: A Structure-aware Text-to-Graph Model for Entity and Relation Extraction*. Apr. 18, 2024. DOI: 10.48550/arXiv.2404.12491. arXiv: 2404.12491[cs]. URL: http://arxiv.org/abs/2404.12491 (cit. on p. 15).

[107] Jinwoo Kim, Dat Tien Nguyen, Seonwoo Min, Sungjun Cho, Moontae Lee, Honglak Lee, and Seunghoon Hong. “Pure Transformers are Powerful Graph Learners”. In: Advances in Neural Information Processing Systems. Oct. 31, 2022. URL: https://openreview.net/forum?id=um2BxfgkT2_ (cit. on p. 15).

[108] Tianyang Zhao, Zhao Yan, Yunbo Cao, and Zhoujun Li. “Asking Effective and Diverse Questions: A Machine Reading Comprehension based Framework for Joint Entity-Relation Extraction”. In: Twenty-Ninth International Joint Conference on Artificial Intelligence. Vol. 4. ISSN: 1045-0823. July 9, 2020, pp. 3948–3954. DOI: 10.24963/ijcai.2020/546. URL: https://www.ijcai.org/proceedings/2020/546 (cit. on p. 15).

[109] Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. “Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples!” In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Findings 2023. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 10572–10601. DOI: 10.18653/v1/2023.findings-emnlp.710. URL: https://aclanthology.org/2023.findings-emnlp.710 (cit. on pp. 15, 16, 19).

[110] Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. *LLM-DA: Data Augmentation via Large Language Models for Few-Shot Named Entity Recognition*. Feb. 22, 2024. DOI: 10.48550/arXiv.2402.14568. arXiv: 2402.14568[cs]. URL: http://arxiv.org/abs/2402.14568 (cit. on pp. 15, 16).

[111] Markus Eberts and Adrian Ulges. “Span-Based Joint Entity and Relation Extraction with Transformer Pre-Training”. In: *ECAI 2020*. IOS Press, 2020, pp. 2006–2013. DOI: 10.3233/FAIA200321. URL: https://ebooks.iospress.nl/doi/10.3233/FAIA200321 (cit. on p. 15).

[112] Yijun Wang, Changzhi Sun, Yuanbin Wu, Hao Zhou, Lei Li, and Junchi Yan. “ENPAR: Enhancing entity and entity pair representations for joint entity relation extraction”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2021, pp. 2877–2887. ISBN: 978-1-954085-02-2. DOI: 10.18653/v1/2021.eacl-main.251. URL: https://aclanthology.org/2021.eacl-main.251 (cit. on pp. 15, 16).

[113] Zexuan Zhong and Danqi Chen. “A Frustratingly Easy Approach for Entity and Relation Extraction”. In: *Proceedings of the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2021, pp. 50–61. DOI: 10.18653/v1/2021.naacl-main.5 (cit. on pp. 15, 45, 51, 61, 77, 79, 83, 97).

[114] Yijun Wang, Changzhi Sun, Yuanbin Wu, Hao Zhou, Lei Li, and Junchi Yan. “UniRE: A Unified Label Space for Entity Relation Extraction”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. ACL-IJCNLP 2021. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, Aug. 2021, pp. 220–231. DOI:

10.18653/v1/2021.acl-long.19. URL: https://aclanthology.org/2021.acl-long.19 (cit. on p. 15).

[115] Youmi Ma, Tatsuya Hiraoka, and Naoaki Okazaki. "Named Entity Recognition and Relation Extraction Using Enhanced Table Filling by Contextualized Representations". In: *Journal of Natural Language Processing* 29.1 (2022), pp. 187–223. DOI: 10.5715/jnlp.29.187 (cit. on pp. 15, 21).

[116] Jianing Wang, Chengyu Wang, Chuanqi Tan, Minghui Qiu, Songfang Huang, Jun Huang, and Ming Gao. "SpanProto: A Two-stage Span-based Prototypical Network for Few-shot Named Entity Recognition". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3466–3476. DOI: 10.18653/v1/2022.emnlp-main.227. URL: https://aclanthology.org/2022.emnlp-main.227 (cit. on pp. 15, 77, 78, 79, 83, 97).

[117] Youmi Ma, Tatsuya Hiraoka, and Naoaki Okazaki. "Joint Entity and Relation Extraction Based on Table Labeling Using Convolutional Neural Networks". In: *Proceedings of the Sixth Workshop on Structured Prediction for NLP*. spnlp 2022. Ed. by Andreas Vlachos, Priyanka Agrawal, André Martins, Gerasimos Lampouras, and Chunchuan Lyu. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 11–21. DOI: 10.18653/v1/2022.spnlp-1.2. URL: https://aclanthology.org/2022.spnlp-1.2 (cit. on p. 15).

[118] Hang Yan, Yu Sun, Xiaonan Li, Yunhua Zhou, Xuanjing Huang, and Xipeng Qiu. "UTC-IE: A Unified Token-pair Classification Architecture for Information Extraction". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2023. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 4096–4122. DOI: 10.18653/v1/2023.acl-long.226. URL: https://aclanthology.org/2023.acl-long.226 (cit. on p. 15).

[119] Debajyoti Chatterjee. *Making Neural Machine Reading Comprehension Faster*. Mar. 29, 2019. DOI: 10.48550/arXiv.1904.00796. arXiv: 1904.00796[cs]. URL: http://arxiv.org/abs/1904.00796 (cit. on p. 15).

[120] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. *ERNIE: Enhanced Representation through Knowledge Integration*. Apr. 19, 2019. DOI: 10.48550/arXiv.1904.09223. arXiv: 1904.09223[cs]. URL: http://arxiv.org/abs/1904.09223 (cit. on pp. 15, 104).

[121] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. "Matching the blanks: Distributional similarity for relation learning". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 2895–2905. ISBN: 978-1-950737-48-2. DOI: 10.18653/v1/p19-1279. URL: https://aclanthology.org/P19-1279 (cit. on pp. 16, 56).

[122] Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne Bernard. *NuNER: Entity Recognition Encoder Pre-training via LLM-Annotated Data*. Feb. 23, 2024. DOI: 10.48550/arXiv.2402.15343. arXiv: 2402.15343[cs]. URL: http://arxiv.org/abs/2402.15343 (cit. on pp. 16, 18).

[123] Letian Peng, Zilong Wang, Feng Yao, Zihan Wang, and Jingbo Shang. *MetaIE: Distilling a Meta Model from LLM for All Kinds of Information Extraction Tasks*. Mar. 30, 2024. DOI: 10.48550/arXiv.2404.00457. arXiv: 2404.00457[cs]. URL: http://arxiv.org/abs/2404.00457 (cit. on p. 16).

[124] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov,

Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. *GPT-4 Technical Report*. Mar. 4, 2024. DOI: 10.48550/arXiv.2303.08774. arXiv: 2303.08774[cs]. URL: http://arxiv.org/abs/2303.08774 (cit. on pp. 16, 17, 78).

[125] Yuji Naraki, Ryosuke Yamaki, Yoshikazu Ikeda, Takafumi Horie, and Hiroki Naganuma. *Augmenting NER Datasets with LLMs: Towards Automated and Refined Annotation*. Mar. 30, 2024. DOI: 10.48550/arXiv.2404.01334. arXiv: 2404.01334[cs]. URL: http://arxiv.org/abs/2404.01334 (cit. on p. 16).

[126] Lilong Xue, Dan Zhang, Yuxiao Dong, and Jie Tang. *AutoRE: Document-Level Relation Extraction with Large Language Models*. July 26, 2024. DOI: 10.48550/arXiv.2403.14888. arXiv: 2403.14888[cs]. URL: http://arxiv.org/abs/2403.14888 (cit. on pp. 16, 18, 20, 21, 50, 53).

[127] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to Sequence Learning with Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html (cit. on p. 16).

[128] Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Mausam, and Soumen Chakrabarti. "IMoJIE: Iterative Memory-Based Joint Open Information Extraction". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 5871–5886. DOI: 10.18653/v1/2020.acl-main.521. URL: https://aclanthology.org/2020.acl-main.521 (cit. on pp. 16, 22).

[129] Pere-Lluís Huguet Cabot and Roberto Navigli. "REBEL: Relation Extraction By End-to-end Language generation". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Findings 2021. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Punta Cana, Dominican Republic: Association for

Computational Linguistics, Nov. 2021, pp. 2370–2381. DOI: 10.18653/v1/2021.findings-emnlp.204. URL: https://aclanthology.org/2021.findings-emnlp.204 (cit. on p. 16).

[130] Martin Josifoski, Nicola De Cao, Maxime Peyrard, Fabio Petroni, and Robert West. "GenIE: Generative Information Extraction". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2022. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 4626–4643. DOI: 10.18653/v1/2022.naacl-main.342. URL: https://aclanthology.org/2022.naacl-main.342 (cit. on p. 16).

[131] Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. "Structured Prediction as Translation between Augmented Natural Languages". In: International Conference on Learning Representations. 2021. URL: https://openreview.net/forum?id=US-TP-xnXI (cit. on p. 16).

[132] Saul B. Needleman and Christian D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". In: *Journal of Molecular Biology* 48.3 (Mar. 28, 1970), pp. 443–453. ISSN: 0022-2836. DOI: 10.1016/0022-2836(70)90057-4. URL: https://www.sciencedirect.com/science/article/pii/0022283670900574 (cit. on pp. 16, 33, 35, 111).

[133] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67. ISSN: 1533-7928. URL: http://jmlr.org/papers/v21/20-074.html (cit. on pp. 17, 87).

[134] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. "Scaling Instruction-Finetuned Language Models". In: *Journal of Machine Learning Research* 25.70 (2024), pp. 1–53. ISSN: 1533-7928. URL: http://jmlr.org/papers/v25/23-0870.html (cit. on pp. 17, 87).

[135] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703 (cit. on p. 17).

[136] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. “Improving language understanding by generative pre-training”. In: (2018) (cit. on p. 17).

[137] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9 (cit. on p. 17).

[138] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. “Training language models to follow instructions with human feedback”. In: *Advances in Neural Information Processing Systems* 35 (Dec. 6, 2022), pp. 27730–27744. URL: https://proceedings.neurips.cc//paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html (cit. on p. 17).

[139] Yuyang Ding, Juntao Li, Pinzheng Wang, Zecheng Tang, Bowen Yan, and Min Zhang. *Rethinking Negative Instances for Generative Named Entity Recognition*. June 18, 2024. DOI: 10.48550/arXiv.2402.16602. arXiv: 2402.16602[cs]. URL: http://arxiv.org/abs/2402.16602 (cit. on pp. 17, 19, 87, 88).

[140] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. *Evaluating Large Language Models Trained on Code*. July 14, 2021. DOI: 10.48550/arXiv.2107.03374. arXiv: 2107.03374[cs]. URL: http://arxiv.org/abs/2107.03374 (cit. on p. 17).

[141] Qi Sun, Kun Huang, Xiaocui Yang, Rong Tong, Kun Zhang, and Soujanya Poria. “Consistency Guided Knowledge Retrieval and Denoising in LLMs for Zero-shot Document-level Relation Triplet Extraction”. In: *Proceedings of the ACM on Web Conference 2024*. WWW ’24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 4407–4416. DOI: 10.1145/3589334.3645678. URL: https://doi.org/10.1145/3589334.3645678 (cit. on pp. 17, 18, 20, 21, 50).

[142] Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. “LasUIE: Unifying Information Extraction with Latent Adaptive Structure-aware Generative Language Model”. In: *Advances in Neural Information Processing Systems* 35 (Dec. 6, 2022), pp. 15460–15475. URL: https://proceedings.

neurips.cc/paper_files/paper/2022/hash/63943ee9fe347f3d95892cf87d9a42e6-Abstract-Conference.html (cit. on p. 17).

[143] Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. *InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction*. Apr. 17, 2023. DOI: 10.48550/arXiv.2304.08085. arXiv: 2304.08085[cs]. URL: http://arxiv.org/abs/2304.08085 (cit. on pp. 17, 18, 80).

[144] Chen Ling, Xujiang Zhao, Xuchao Zhang, Yanchi Liu, Wei Cheng, Haoyu Wang, Zhengzhang Chen, Takao Osaki, Katsushi Matsuda, Haifeng Chen, and Liang Zhao. *Improving Open Information Extraction with Large Language Models: A Study on Demonstration Uncertainty*. Sept. 6, 2023. DOI: 10.48550/arXiv.2309.03433. arXiv: 2309.03433[cs]. URL: http://arxiv.org/abs/2309.03433 (cit. on pp. 17, 22).

[145] Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. "Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Findings 2023. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 794–812. DOI: 10.18653/v1/2023.findings-acl.50. URL: https://aclanthology.org/2023.findings-acl.50 (cit. on pp. 17, 21).

[146] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. *GPT-NER: Named Entity Recognition via Large Language Models*. May 12, 2023. DOI: 10.48550/arXiv.2304.10428. arXiv: 2304.10428[cs]. URL: http://arxiv.org/abs/2304.10428 (cit. on pp. 17, 80).

[147] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *Advances in Neural Information Processing Systems* 35 (Dec. 6, 2022), pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html (cit. on pp. 17, 18).

[148] Yucan Guo, Zixuan Li, Xiaolong Jin, Yantao Liu, Yutao Zeng, Wenxuan Liu, Xiang Li, Pan Yang, Long Bai, Jiafeng Guo, and Xueqi Cheng. *Retrieval-Augmented Code Generation for Universal Information Extraction*. Nov. 6, 2023. DOI: 10.48550/arXiv.2311.02962. arXiv: 2311.02962[cs]. URL: http://arxiv.org/abs/2311.02962 (cit. on p. 17).

[149] Xingyao Wang, Sha Li, and Heng Ji. "Code4Struct: Code Generation for Few-Shot Event Structure Prediction". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2023. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 3640–3663. DOI: 10.18653/v1/2023.acl-long.202. URL: https://aclanthology.org/2023.acl-long.202 (cit. on p. 17).

[150] Zhen Bi, Jing Chen, Yinuo Jiang, Feiyu Xiong, Wei Guo, Huajun Chen, and Ningyu Zhang. "CodeKGC: Code Language Model for Generative Knowledge Graph Construction". In: *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 23.3 (Mar. 9, 2024), 45:1–45:16. ISSN: 2375-4699. DOI: 10.1145/3641850. URL: https://doi.org/10.1145/3641850 (cit. on p. 18).

[151] Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. *Zero-Shot Information Extraction via Chatting with ChatGPT*. Feb. 20, 2023. DOI: 10.48550/arXiv.2302.10205. arXiv: 2302.10205[cs]. URL: http://arxiv.org/abs/2302.10205 (cit. on pp. 18, 80, 87, 88, 90).

[152] Somin Wadhwa, Silvio Amir, and Byron Wallace. "Revisiting Relation Extraction in the era of Large Language Models". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2023. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 15566–15589. DOI: 10.18653/v1/2023.acl-long.868. URL: https://aclanthology.org/2023.acl-long.868 (cit. on p. 18).

[153] Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. "GPT-RE: In-context Learning for Relation Extraction using Large Language Models". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2023. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 3534–3547. DOI: 10.18653/v1/2023.emnlp-main.214. URL: https://aclanthology.org/2023.emnlp-main.214 (cit. on pp. 18, 19).

[154] Mingchen Li, Huixue Zhou, Han Yang, and Rui Zhang. "RT: a Retrieving and Chain-of-Thought framework for few-shot medical named entity recognition". In: *Journal of the American Medical Informatics Association* (May 6, 2024), ocae095. ISSN: 1527-974X. DOI: 10.1093/jamia/ocae095. URL: https://doi.org/10.1093/jamia/ocae095 (cit. on p. 18).

[155] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (cit. on pp. 18, 57, 78).

[156] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. "Distant supervision for relation extraction without labeled data". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on*

Natural Language. Suntec, Singapore: Association for Computational Linguistics, 2009, pp. 1003–1011. DOI: 10.3115/1690219.1690287 (cit. on pp. 18, 55).

[157] Sebastian Riedel, Limin Yao, and Andrew McCallum. "Modeling relations and their mentions without labeled text". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 6323 LNAI. Issue: PART 3. Berlin, Heidelberg: Springer, 2010, pp. 148–163. ISBN: 3-642-15938-9. DOI: 10.1007/978-3-642-15939-8_10 (cit. on pp. 18, 28, 55).

[158] Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. "Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2023. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1555–1574. DOI: 10.18653/v1/2023.emnlp-main.96. URL: https://aclanthology.org/2023.emnlp-main.96 (cit. on p. 18).

[159] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*. Dec. 31, 2020. DOI: 10.48550/arXiv.2101.00027. arXiv: 2101.00027[cs]. URL: http://arxiv.org/abs/2101.00027 (cit. on pp. 18, 80, 89).

[160] Ishani Mondal, Michelle Yuan, Anandhavelu N, Aparna Garimella, Francis Ferraro, Andrew Blair-Stanek, Benjamin Van Durme, and Jordan Boyd-Graber. *InteractiveIE: Towards Assessing the Strength of Human-AI Collaboration in Improving the Performance of Information Extraction*. May 23, 2023. DOI: 10.48550/arXiv.2305.14659. arXiv: 2305.14659[cs]. URL: http://arxiv.org/abs/2305.14659 (cit. on p. 19).

[161] Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. *Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors*. May 23, 2023. DOI: 10.48550/arXiv.2305.14450. arXiv: 2305.14450[cs]. URL: http://arxiv.org/abs/2305.14450 (cit. on p. 19).

[162] Xavier Carreras and Lluís Màrquez. "Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling". In: *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*. CoNLL-HLT 2004. Boston, Massachusetts, USA: Association for Computational Linguistics, May 6, 2004, pp. 89–97. URL: https://aclanthology.org/W04-2412 (cit. on p. 19).

[163] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. "Language models as knowledge bases?" In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2019, pp. 2463–2473. ISBN: 978-1-950737-90-1. DOI: 10.18653/v1/d19-1250. URL: https://aclanthology.org/D19-1250 (cit. on p. 19).

[164] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. "Large language models encode clinical knowledge". In: *Nature* 620.7972 (Aug. 2023). Publisher: Nature Publishing Group, pp. 172–180. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06291-2. URL: https://www.nature.com/articles/s41586-023-06291-2 (cit. on p. 19).

[165] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html (cit. on p. 19).

[166] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. "Unifying Large Language Models and Knowledge Graphs: A Roadmap". In: *IEEE Transactions on Knowledge and Data Engineering* 36.7 (July 2024). Conference Name: IEEE Transactions on Knowledge and Data Engineering, pp. 3580–3599. ISSN: 1558-2191. DOI: 10.1109/TKDE.2024.3352100. URL: https://ieeexplore.ieee.org/abstract/document/10387715 (cit. on p. 19).

[167] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. "Object Hallucination in Image Captioning". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2018. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 4035–4045. DOI: 10.18653/v1/D18-1437. URL: https://aclanthology.org/D18-1437 (cit. on p. 20).

[168] Yijun Xiao and William Yang Wang. "On Hallucination and Predictive Uncertainty in Conditional Language Generation". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. EACL 2021. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics, Apr. 2021, pp. 2734–2744. DOI: 10.18653/v1/2021.eacl-main.236. URL: https://aclanthology.org/2021.eacl-main.236 (cit. on p. 20).

[169] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. *From Local to Global: A Graph RAG Approach to Query-Focused Summarization*. Apr. 24, 2024. DOI: 10.48550/arXiv.2404.16130. arXiv: 2404.16130[cs]. URL: http://arxiv.org/abs/2404.16130 (cit. on p. 20).

[170] Tyler Procko. *Graph Retrieval-Augmented Generation for Large Language Models: A Survey*. Rochester, NY, July 13, 2024. URL: https://papers.ssrn.com/abstract=4895062 (cit. on p. 20).

[171] Costas Mavromatis and George Karypis. *GNN-RAG: Graph Neural Retrieval for Large Language Model Reasoning*. May 30, 2024. DOI: 10.48550/arXiv.2405.20139. arXiv: 2405.20139[cs]. URL: http://arxiv.org/abs/2405.20139 (cit. on p. 20).

[172] Kumutha Swampillai and Mark Stevenson. "Extracting relations within and across sentences". In: *Proceedings of the international conference recent advances in natural language processing 2011*. Hissar, Bulgaria, 2011, pp. 25–32 (cit. on p. 20).

[173] Liyan Xu and Jinho Choi. "Modeling Task Interactions in Document-Level Joint Entity and Relation Extraction". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2022. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 5409–5416. DOI: 10.18653/v1/2022.naacl-main.395. URL: https://aclanthology.org/2022.naacl-main.395 (cit. on pp. 20, 21).

[174] Wang Xu, Kehai Chen, and Tiejun Zhao. "Document-Level Relation Extraction with Path Reasoning". In: *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 22.4 (Mar. 25, 2023), 104:1–104:14. ISSN: 2375-4699. DOI: 10.1145/3572898. URL: https://doi.org/10.1145/3572898 (cit. on p. 20).

[175] Ruoyu Zhang, Yanzeng Li, and Lei Zou. "A Novel Table-to-Graph Generation Approach for Document-Level Joint Entity and Relation Extraction". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2023. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 10853–10865. DOI: 10.18653/v1/2023.acl-long.607. URL: https://aclanthology.org/2023.acl-long.607 (cit. on pp. 20, 50).

[176] Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. "Double graph based reasoning for document-level relation extraction". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 2020 Conference on Empirical Methods in Natural Language Processing. Online: Association for Computational Linguistics (ACL), Sept. 2020, pp. 1630–1640. ISBN: 978-1-952148-60-6. DOI: 10.18653/v1/2020.emnlp-main.127. URL: https://arxiv.org/abs/2009.13752v1 (cit. on pp. 20, 28, 50, 51, 112).

[177] Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Tingwen Liu, Hengzhu Tang, Wang Yubin, and Li Guo. "Document-level Relation Extraction with Dual-tier Heterogeneous Graph". In: *Proceedings of the 28th International Conference on Computational Linguistics*. 28th International Conference on Computational Linguistics. Barcelona, Spain: Association for Computational Linguistics, Jan. 2020, pp. 1630–1641. DOI: 10.18653/v1/2020.coling-main.143. URL: https://aclanthology.org/2020.coling-main.143 (cit. on pp. 20, 28, 51).

[178] Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. "Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Conference on Artificial Intelligence. Vol. 35. Number: 16. Online: AAAI Press, May 18, 2021, pp. 14612–

14620. DOI: 10.1609/aaai.v35i16.17717. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17717 (cit. on pp. 21, 50, 51, 52, 112).

[179] Xiusheng Huang, Hang Yang, Yubo Chen, Jun Zhao, Kang Liu, Weijian Sun, and Zuyu Zhao. "Document-Level Relation Extraction via Pair-Aware and Entity-Enhanced Representation Learning". In: *Proceedings of the 29th International Conference on Computational Linguistics*. COLING 2022. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 2418–2428. URL: https://aclanthology.org/2022.coling-1.213 (cit. on p. 21).

[180] Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. "Three Sentences Are All You Need: Local Path Enhanced Document Relation Extraction". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Vol. 2. Online: Association for Computational Linguistics, June 2021, pp. 998–1004. ISBN: 978-1-954085-52-7. DOI: 10.18653/v1/2021.acl-short.126. URL: https://arxiv.org/abs/2106.01793v1 (cit. on pp. 21, 28, 50).

[181] Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. "Eider: Empowering Document-level Relation Extraction with Efficient Evidence Extraction and Inference-stage Fusion". In: *Findings of the Association for Computational Linguistics: ACL 2022*. Findings 2022. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 257–268. DOI: 10.18653/v1/2022.findings-acl.23. URL: https://aclanthology.org/2022.findings-acl.23 (cit. on p. 21).

[182] Youmi Ma, An Wang, and Naoaki Okazaki. "DREEAM: Guiding Attention with Evidence for Improving Document-Level Relation Extraction". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. EACL 2023. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1971–1983. DOI: 10.18653/v1/2023.eacl-main.145. URL: https://aclanthology.org/2023.eacl-main.145 (cit. on pp. 21, 112).

[183] Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, and Jiawei Han. "SAIS: Supervising and Augmenting Intermediate Steps for Document-Level Relation Extraction". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2022. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 2395–2409. DOI: 10.18653/v1/2022.naacl-main.171. URL: https://aclanthology.org/2022.naacl-main.171 (cit. on p. 21).

[184] Junpeng Li, Zixia Jia, and Zilong Zheng. “Semi-automatic Data Enhancement for Document-Level Relation Extraction with Distant Supervision from Large Language Models”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2023. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 5495–5505. DOI: 10.18653/v1/2023.emnlp-main.334. URL: https://aclanthology.org/2023.emnlp-main.334 (cit. on pp. 21, 112).

[185] Bill MacCartney. *Natural language inference*. Stanford University, 2009 (cit. on p. 21).

[186] Xinyi Wang, Zitao Wang, Weijian Sun, and Wei Hu. “Enhancing Document-Level Relation Extraction by Entity Knowledge Injection”. In: *The Semantic Web – ISWC 2022*. Ed. by Ulrike Sattler, Aidan Hogan, Maria Keet, Valentina Presutti, João Paulo A. Almeida, Hideaki Takeda, Pierre Monnin, Giuseppe Pirrò, and Claudia d’Amato. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022, pp. 39–56. ISBN: 978-3-031-19433-7. DOI: 10.1007/978-3-031-19433-7_3 (cit. on pp. 21, 50).

[187] Kuicai Dong, Zhao Yilin, Aixin Sun, Jung-Jae Kim, and Xiaoli Li. “DocOIE: A Document-level Context-Aware Dataset for OpenIE”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Findings 2021. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, Aug. 2021, pp. 2377–2389. DOI: 10.18653/v1/2021.findings-acl.210. URL: https://aclanthology.org/2021.findings-acl.210 (cit. on pp. 21, 22).

[188] Markus Eberts and Adrian Ulges. “An End-to-end Model for Entity-level Relation Extraction using Multi-instance Learning”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. EACL 2021. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics, Apr. 2021, pp. 3650–3660. DOI: 10.18653/v1/2021.eacl-main.319. URL: https://aclanthology.org/2021.eacl-main.319 (cit. on p. 21).

[189] Arpita Roy, Youngja Park, Taesung Lee, and Shimei Pan. “Supervising unsupervised open information extraction models”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 728–737. ISBN: 978-1-950737-90-1. DOI: 10.18653/v1/d19-1067 (cit. on p. 22).

[190] Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. “OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. EMNLP 2020. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 3748–3761. DOI: 10.18653/v1/2020.emnlp-main.306. URL: https://aclanthology.org/2020.emnlp-main.306 (cit. on p. 22).

[191] Dinesh Nagumothu, Bahadorreza Ofoghi, Guangyan Huang, and Peter Eklund. "PIE-QG: Paraphrased Information Extraction for Unsupervised Question Generation from Small Corpora". In: *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*. CoNLL 2022. Ed. by Antske Fokkens and Vivek Srikumar. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 350–359. DOI: 10.18653/v1/2022.conll-1.24. URL: https://aclanthology.org/2022.conll-1.24 (cit. on p. 22).

[192] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization". In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. PMLR, Nov. 21, 2020, pp. 11328–11339. URL: https://proceedings.mlr.press/v119/zhang20ae.html (cit. on p. 22).

[193] Bowen Yu, Yucheng Wang, Tingwen Liu, Hongsong Zhu, Limin Sun, and Bin Wang. "Maximal Clique Based Non-Autoregressive Open Information Extraction". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2021. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 9696–9706. DOI: 10.18653/v1/2021.emnlp-main.764. URL: https://aclanthology.org/2021.emnlp-main.764 (cit. on p. 22).

[194] Youngbin Ro, Yukyung Lee, and Pilsung Kang. "Multi^2OIE: Multilingual Open Information Extraction Based on Multi-Head Attention with BERT". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Findings 2020. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 1107–1117. DOI: 10.18653/v1/2020.findings-emnlp.99. URL: https://aclanthology.org/2020.findings-emnlp.99 (cit. on p. 23).

[195] Kuicai Dong, Aixin Sun, Jung-Jae Kim, and Xiaoli Li. "Syntactic Multi-view Learning for Open Information Extraction". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2022. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 4072–4083. DOI: 10.18653/v1/2022.emnlp-main.272. URL: https://aclanthology.org/2022.emnlp-main.272 (cit. on p. 23).

[196] Ruidong Wu, Yuan Yao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. "Open relation extraction: Relational knowledge transfer from supervised data to unsupervised data". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 219–228. ISBN: 978-1-950737-90-1. DOI: 10.18653/v1/d19-1021 (cit. on pp. 23, 55, 56, 58, 59, 61).

[197] Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip S. Yu. "SelfORE: Self-supervised relational feature learning for open relation extraction". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Online:

Association for Computational Linguistics, 2020, pp. 3673–3682. ISBN: 978-1-952148-60-6. DOI: 10.18653/v1/2020.emnlp-main.299 (cit. on pp. 23, 55, 56, 57, 58, 59, 61, 63, 65, 110).

[198] Jun Zhao, Tao Gui, Qi Zhang, and Yaqian Zhou. "A Relation-Oriented Clustering Method for Open Relation Extraction". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 9707–9718 (cit. on pp. 23, 56, 57, 58, 61, 63, 66, 112).

[199] Bin Duan, Shusen Wang, Xingxian Liu, and Yajing Xu. "Cluster-aware Pseudo-Labeling for Supervised Open Relation Extraction". In: *Proceedings of the 29th International Conference on Computational Linguistics*. COLING 2022. Ed. by Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 1834–1841. URL: https://aclanthology.org/2022.coling-1.158 (cit. on pp. 23, 112).

[200] Xuming Hu, Zhaochen Hong, Chenwei Zhang, Aiwei Liu, Shiao Meng, Lijie Wen, Irwin King, and Philip S. Yu. "Reading Broadly to Open Your Mind: Improving Open Relation Extraction With Search Documents Under Self-Supervisions". In: *IEEE Transactions on Knowledge and Data Engineering* 36.5 (Sept. 2023). Conference Name: IEEE Transactions on Knowledge and Data Engineering, pp. 2026–2040. ISSN: 1558-2191. DOI: 10.1109/TKDE.2023.3317139. URL: https://ieeexplore.ieee.org/abstract/document/10255305 (cit. on p. 23).

[201] Jiaxin Wang, Lingling Zhang, Jun Liu, Xi Liang, Yujie Zhong, and Yaqiang Wu. "MatchPrompt: Prompt-based Open Relation Extraction with Semantic Consistency Guided Clustering". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2022. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 7875–7888. DOI: 10.18653/v1/2022.emnlp-main.537. URL: https://aclanthology.org/2022.emnlp-main.537 (cit. on pp. 23, 66).

[202] Jun Zhao, Yongxin Zhang, Qi Zhang, Tao Gui, Zhongyu Wei, Minlong Peng, and Mingming Sun. "Actively Supervised Clustering for Open Relation Extraction". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2023. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 4985–4997. DOI: 10.18653/v1/2023.acl-long.273. URL: https://aclanthology.org/2023.acl-long.273 (cit. on pp. 23, 66, 112, 113).

[203] Filipe Mesquita, Matteo Cannaviccio, Jordan Schmidek, Paramita Mirza, and Denilson Barbosa. "Knowledgenet: A benchmark dataset for knowledge base population". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing and 9th International Joint Conference on Natural Language Processing. 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 749–758. ISBN: 978-1-950737-90-1. DOI: 10.18653/v1/d19-1069. URL: https://aclanthology.org/D19-1069 (cit. on pp. 26, 27, 28, 29, 30).

[204] Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. "HacRED: A Large-Scale Relation Extraction Dataset Toward Hard Cases in Practical Applications". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online: Association for Computational Linguistics, 2021, pp. 2819–2831. DOI: 10.18653/v1/2021.findings-acl.249. URL: https://aclanthology.org/2021.findings-acl.249 (cit. on pp. 26, 27, 28, 29, 30).

[205] Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Elena Simperl, and Frederique Laforest. "T-Rex: A large scale alignment of natural language with knowledge base triples". In: *Proceedings of the 11th International Conference on Language Resources and Evaluation*. 11th International Conference on Language Resources. Miyazaki, Japan: European Language Resources Association, 2018, pp. 3448–3452. ISBN: 979-10-95546-00-9. URL: https://aclanthology.org/L18-1544 (cit. on pp. 27, 28, 29, 30).

[206] Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. "Fewrel 2.0: Towards more challenging few-shot relation classification". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 6250–6255. ISBN: 978-1-950737-90-1. DOI: 10.18653/v1/d19-1649. URL: https://aclanthology.org/D19-1649 (cit. on pp. 27, 28, 29, 30, 57, 65).

[207] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. "Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2018. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 3219–3232. DOI: 10.18653/v1/D18-1360. URL: https://aclanthology.org/D18-1360 (cit. on p. 28).

[208] Arpita Roy and Shimei Pan. "Incorporating medical knowledge in BERT for clinical relation extraction". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: Association for Computational Linguistics, Dec. 2021, pp. 5357–5366. DOI: 10.18653/v1/2021.emnlp-main.435. URL: https://aclanthology.org/2021.emnlp-main.435 (cit. on p. 28).

[209] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. "BioCreative V CDR task corpus: a resource for chemical disease relation extraction". In: *Database* 2016 (Jan. 1, 2016), baw068. ISSN: 1758-0463. DOI: 10.1093/database/baw068. URL: https://doi.org/10.1093/database/baw068 (cit. on pp. 28, 29, 30, 78).

[210] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. "Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 4803–4809. ISBN: 978-1-948087-84-1. DOI: 10.18653/v1/d18-1514 (cit. on pp. 28, 29, 30, 64).

[211] Pablo Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. "DBpedia spotlight: Shedding light on the web of documents". In: *Proceedings of the 7th International Conference on Semantic Systems*. 7th International Conference on Semantic Systems. Graz, Austria: Association for Computing Machinery, 2011, pp. 1–8. ISBN: 978-1-4503-0621-8. DOI: 10.1145/2063518.2063519 (cit. on p. 30).

[212] Stephen Robertson, S Walker, S Jones, M. M. Hancock-Beaulieu, and M. Gatford. "Okapi at TREC-3". In: *Overview of the Third Text REtrieval Conference (TREC-3)*. The third text REtrieval conference (TREC-3). Gaithersburg, MD, USA: DIANE Publishing Company, 1994, pp. 109–126 (cit. on p. 33).

[213] V. I. Levenshtein. "Binary codes capable of correcting deletions, insertions and reversals". In: *Soviet Physics Doklady* 10 (Feb. 1966), p. 707 (cit. on p. 33).

[214] Edwin B. Wilson. "Probable inference, the law of succession, and statistical inference". In: *Journal of the American Statistical Association* 22.158 (1927), pp. 209–212. DOI: 10.1080/01621459.1927.10502953. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1927.10502953 (cit. on p. 34).

[215] Jacob Cohen. "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46 (cit. on p. 40).

[216] Paolo Ferragina and Ugo Scaiella. "Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)". In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. 19th ACM international conference on Information and knowledge management. Toronto, ON, Canada: Association for Computing Machinery, 2010, pp. 1625–1628 (cit. on p. 41).

[217] Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. "Revisiting DocRED - Addressing the False Negative Problem in Relation Extraction". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 8472–8487. URL: https://aclanthology.org/2022.emnlp-main.580 (cit. on p. 44).

[218] Amit Bagga and Breck Baldwin. "Algorithms for scoring coreference chain". In: *Proceedings of the 1st International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*. First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference. Granada, Spain: European Language Resources Association, 1998, pp. 563–566 (cit. on pp. 46, 49, 58).

[219] Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. "Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation". In: *Proceedings of the conference. Association for Computational Linguistics. Meeting* 2014 (June 2014), pp. 30–35. ISSN: 0736-587X. DOI: 10.3115/v1/P14-2006. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5667668/ (cit. on p. 46).

[220] Nafise Sadat Moosavi and Michael Strube. "Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2016. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 632–642. DOI: 10.18653/v1/P16-1060. URL: https://aclanthology.org/P16-1060 (cit. on p. 46).

[221] Andrew Rosenberg and Julia Hirschberg. "V-Measure: A conditional entropy-based external cluster evaluation measure". In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 410–420 (cit. on pp. 49, 58).

[222] Nguyen Xuan Vinh, Julien Epps, and James Bailey. "Information theoretic measures for clusterings comparison: is a correction for chance necessary?" In: *Proceedings of the 26th annual international conference on machine learning*. 2009, pp. 1073–1080 (cit. on p. 49).

[223] Etienne Simon, Vincent Guigue, and Benjamin Piwowarski. "Unsupervised information extraction: Regularizing discriminative approaches with relation distribution losses". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 1378–1387. ISBN: 978-1-950737-48-2. DOI: 10.18653/v1/p19-1133 (cit. on pp. 49, 57, 58, 59, 65, 66).

[224] Simone Romano, Nguyen Xuan Vinh, James Bailey, and Karin Verspoor. "Adjusting for chance clustering comparison measures". In: *The Journal of Machine Learning Research* 17.1 (Jan. 1, 2016), pp. 4635–4666. ISSN: 1532-4435. URL: https://dl.acm.org/doi/abs/10.5555/2946645.3007087 (cit. on p. 49).

[225] Iz Beltagy, Matthew E. Peters, and Arman Cohan. *Longformer: The Long-Document Transformer*. Dec. 2, 2020. DOI: 10.48550/arXiv.2004.05150. arXiv: 2004.05150[cs]. URL: http://arxiv.org/abs/2004.05150 (cit. on p. 51).

[226] Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. "Long-tail relation extraction via knowledge graph embeddings and graph convolution networks". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019, pp. 3016–3025. ISBN: 978-1-950737-13-0. DOI: 10.18653/v1/n19-1306. URL: https://aclanthology.org/N19-1306 (cit. on p. 51).

[227] Shaoxiong Ji, Shirui Pan, Erik Cambria, Senior Member, Pekka Marttinen, Philip S. Yu, and Life Fellow. "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications". In: *IEEE Transactions on Neural Networks and Learning Systems* (2021). Publisher: Institute of Electrical and Electronics Engineers Inc., pp. 1–21. DOI: 10.1109/TNNLS.2021.3070843 (cit. on p. 55).

[228] Renze Lou, Fan Zhang, Xiaowei Zhou, Yutong Wang, Minghui Wu, and Lin Sun. "A Unified Representation Learning Strategy for Open Relation Extraction with Ranked List Loss". In: *Proceedings of the 20th China National Conference on Computational Linguistics*. Vol. 12869 LNAI. Huhhot, China: Chinese Information Processing Society of China, 2021, pp. 1096–1108. ISBN: 978-3-030-84185-0. DOI: 10.1007/978-3-030-84186-7_21 (cit. on pp. 55, 58, 59, 61, 66).

[229] Bo Lv, Li Jin, Yanan Zhang, Hao Wang, Xiaoyu Li, and Zhi Guo. "Commonsense Knowledge-Aware Prompt Tuning for Few-Shot NOTA Relation Classification". In: *Applied Sciences* 12.4 (Feb. 2022). Publisher: Multidisciplinary Digital Publishing Institute, pp. 2185–2185. DOI: 10.3390/app12042185. URL: https://www.mdpi.com/2076-3417/12/4/2185/htm (cit. on pp. 56, 57, 62, 63, 69, 70).

[230] Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. "KnowPrompt: Knowledge-aware Prompt-tuning with Synergistic Optimization for Relation Extraction". In: *Proceedings of the ACM Web Conference 2022*. Lyon, France: Association for Computing Machinery, Apr. 2022. ISBN: 978-1-4503-9096-5. DOI: 10.48550/arxiv.2104.07650. URL: https://arxiv.org/abs/2104.07650v6 (cit. on pp. 56, 57, 62).

[231] Jiaying Gong and Hoda Eldardiry. *Prompt-based Zero-shot Relation Classification with Semantic Knowledge Augmentation*. Dec. 2021. DOI: 10.48550/arxiv.2112.04539. URL: https://arxiv.org/abs/2112.04539v1 (cit. on pp. 56, 57, 62).

[232] Jiejun Tan, Wenbin Hu, and WeiWei Liu. *EPPAC: Entity Pre-typing Relation Classification with Prompt Answer Centralizing*. Mar. 2022. DOI: 10.48550/arxiv.2203.00193. URL: https://arxiv.org/abs/2203.00193v2 (cit. on pp. 56, 57, 62).

[233] Xu Han, Tianyu Gao, Yuan Yao, Demin Ye, Zhiyuan Liu, and Maosong Sun. "OpenNRE: An open and extensible toolkit for neural relation extraction". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing: System Demonstrations*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 169–174. ISBN: 978-1-950737-92-5. DOI: 10.18653/v1/d19-3029 (cit. on p. 56).

[234] Xu Han, Yi Dai, Tianyu Gao, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. "Continual Relation Learning via Episodic Memory Activation and Reconsolidation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020, pp. 6429–6440. DOI: 10.18653/v1/2020.acl-main.573. URL: https://aclanthology.org/2020.acl-main.573 (cit. on p. 56).

[235] Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. "More Data, More Relations, More Context and More Openness: A Review and Outlook for Relation Extraction". In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, 2020, pp. 745–758 (cit. on p. 56).

[236] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. *Deep contextualized word representations*. Mar. 22, 2018. DOI: 10.48550/arXiv.1802.05365. arXiv: 1802.05365[cs]. URL: http://arxiv.org/abs/1802.05365 (cit. on p. 57).

[237] Jake Snell, Kevin Swersky, and Richard Zemel. "Prototypical networks for few-shot learning". In: *Advances in Neural Information Processing Systems*. Vol. 2017-Decem. Neural information processing systems foundation, Mar. 2017, pp. 4078–4088. DOI: 10.48550/arxiv.1703.05175. URL: https://arxiv.org/abs/1703.05175v2 (cit. on pp. 57, 79).

[238] Haopeng Ren, Yi Cai, Xiaofeng Chen, Guohua Wang, and Qing Li. "A Two-phase Prototypical Network Model for Incremental Few-shot Relation Classification". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain: Association for Computational Linguistics, Jan. 2020, pp. 1618–1629. DOI: 10.18653/v1/2020.coling-main.142. URL: https://aclanthology.org/2020.coling-main.142 (cit. on p. 57).

[239] Tianyu Gao, Adam Fisch, and Danqi Chen. "Making pre-trained language models better few-shot learners". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Online: Association for Computational Linguistics, Dec. 2021, pp. 3816–3830. ISBN: 978-1-954085-52-7. DOI: 10.18653/v1/2021.acl-long.295. URL: https://arxiv.org/abs/2012.15723v2 (cit. on p. 57).

[240] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. July 2021. DOI: 10.48550/arxiv.2107.13586. URL: https://arxiv.org/abs/2107.13586v1 (cit. on p. 57).

[241] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. "How can we know what language models know?" In: *Transactions of the Association for Computational*

Linguistics 8 (2020). Publisher: MIT Press Journals, pp. 423–438. DOI: 10.1162/tacl_a_00324. URL: https://aclanthology.org/2020.tacl-1.28 (cit. on pp. 57, 63, 70).

[242] Derek Tam, Rakesh R. Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. "Improving and Simplifying Pattern Exploiting Training". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2021. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4980–4991. DOI: 10.18653/v1/2021.emnlp-main.407. URL: https://aclanthology.org/2021.emnlp-main.407 (cit. on p. 57).

[243] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. "Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2022. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8086–8098. DOI: 10.18653/v1/2022.acl-long.556. URL: https://aclanthology.org/2022.acl-long.556 (cit. on p. 57).

[244] Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. *Entailment as Few-Shot Learner*. Apr. 29, 2021. DOI: 10.48550/arXiv.2104.14690. arXiv: 2104.14690[cs]. URL: http://arxiv.org/abs/2104.14690 (cit. on p. 57).

[245] Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. "Structured relation discovery using generative models". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK: Association for Computational Linguistics, 2011, pp. 1456–1466. ISBN: 1-937284-11-5 (cit. on p. 58).

[246] David M Blei, Andrew Y Ng, and Michael I. Jordan. "Latent Dirichlet allocation". In: *Journal of Machine Learning Research* 3.4 (2003), pp. 993–1022. DOI: 10.1016/b978-0-12-411519-4.00006-9 (cit. on p. 58).

[247] Limin Yao, Sebastian Riedel, and Andrew McCallum. "Unsupervised relation discovery with sense disambiguation". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Vol. 1. Jeju Island, Korea: Association for Computational Linguistics, 2012, pp. 712–720. ISBN: 978-1-937284-24-4 (cit. on p. 58).

[248] Diego Marcheggiani and Ivan Titov. "Discrete-State Variational Autoencoders for Joint Discovery and Factorization of Relations". In: *Transactions of the Association for Computational Linguistics* 4 (Dec. 2016). Publisher: MIT Press - Journals, pp. 231–244. DOI: 10.1162/tacl_a_00095 (cit. on pp. 58, 59).

[249] Hady Elsahar, Elena Demidova, Simon Gottschalk, Christophe Gravier, and Frederique Laforest. "Unsupervised Open Relation Extraction". In: *The Semantic Web: ESWC 2017 Satellite Events*. Vol. 10577 LNCS. Cham: Springer, Jan. 2017, pp. 12–16. ISBN: 978-3-319-70406-7. DOI: 10.1007/978-3-319-70407-4_3 (cit. on pp. 58, 59, 63).

[250] Diederik P. Kingma and Max Welling. "Auto-encoding variational bayes". In: *Proceedings of the 2nd International Conference on Learning Representations*. Banff, Canada: International Conference on Learning Representations, Dec. 2014. URL: https://arxiv.org/abs/1312.6114v10 (cit. on p. 58).

[251] Chenhan Yuan and Hoda Eldardiry. "Unsupervised Relation Extraction: A Variational Autoencoder Approach". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 1929–1938. DOI: 10.18653/v1/2021.emnlp-main.147. URL: https://aclanthology.org/2021.emnlp-main.147 (cit. on pp. 58, 59, 61, 66).

[252] Thy Thy Tran, Phong Le, and Sophia Ananiadou. "Revisiting Unsupervised Relation Extraction". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7498–7505. DOI: 10.18653/v1/2020.acl-main.669 (cit. on pp. 58, 59, 65, 66).

[253] Sumit Chopra, Raia Hadsell, and Yann LeCun. "Learning a similarity metric discriminatively, with application to face verification". In: *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. I. San Diego, CA, USA: IEEE Computer Society, 2005, pp. 539–546. ISBN: 0-7695-2372-2. DOI: 10.1109/CVPR.2005.202 (cit. on p. 58).

[254] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, and Neil M. Robertson. "Ranked list loss for deep metric learning". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2019-June. IEEE Computer Society, Mar. 2019, pp. 5202–5211. ISBN: 978-1-72813-293-8. DOI: 10.1109/CVPR.2019.00535. URL: https://arxiv.org/abs/1903.03238v8 (cit. on p. 58).

[255] Fangchao Liu, Lingyong Yan, Hongyu Lin, Xianpei Han, and Le Sun. "Element intervention for open relation extraction". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Online: Association for Computational Linguistics, June 2021, pp. 4683–4693. ISBN: 978-1-954085-52-7. DOI: 10.18653/v1/2021.acl-long.361. URL: https://arxiv.org/abs/2106.09558v1 (cit. on p. 61).

[256] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. July 2019. DOI: 10.48550/arxiv.1907.11692. URL: https://arxiv.org/abs/1907.11692v1 (cit. on pp. 61, 104).

[257] Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. *A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models*. Dec. 23, 2023. DOI: 10.48550/arXiv.2303.10420. arXiv: 2303.10420[cs]. URL: http://arxiv.org/abs/2303.10420 (cit. on pp. 62, 80).

[258] Stuart Lloyd. "Least squares quantization in PCM". In: *Technical Report RR-5497* (1957) (cit. on pp. 64, 84, 111).

[259] James MacQueen. “Some methods for classification and analysis of multivariate observations”. In: *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*. Berkeley, California, United States: University of California Press, 1967, pp. 281–297 (cit. on pp. 64, 84, 111).

[260] R. Sibson. “SLINK: An optimally efficient algorithm for the single-link cluster method”. In: *The Computer Journal* 16.1 (Jan. 1973). Publisher: Oxford Academic, pp. 30–34. DOI: 10.1093/comjnl/16.1.30. URL: https://academic.oup.com/comjnl/article/16/1/30/434805 (cit. on p. 64).

[261] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. Portland, Oregon, United States: AAAI Press, 1996, pp. 226–231. URL: www.aaai.org (cit. on p. 64).

[262] Mihael Ankerst, Markus M. Breunig, Hans Peter Kriegel, and Jörg Sander. “OPTICS: Ordering Points to Identify the Clustering Structure”. In: *SIGMOD Record (ACM Special Interest Group on Management of Data)* 28.2 (June 1999). Publisher: ACM PUB27 New York, NY, USA, pp. 49–60. DOI: 10.1145/304181.304187. URL: https://dl.acm.org/doi/abs/10.1145/304181.304187 (cit. on p. 64).

[263] Carl Rasmussen. “The Infinite Gaussian Mixture Model”. In: *Advances in Neural Information Processing Systems*. Vol. 12. MIT Press, 1999. URL: https://proceedings.neurips.cc/paper/1999/hash/97d98119037c5b8a9663cb21fb8ebf47-Abstract.html (cit. on p. 64).

[264] Brendan J. Frey and Delbert Dueck. “Clustering by passing messages between data points”. In: *Science* 315.5814 (Feb. 2007). Publisher: American Association for the Advancement of Science, pp. 972–976. DOI: 10.1126/science.1136800. URL: https://www.science.org/doi/abs/10.1126/science.1136800 (cit. on p. 64).

[265] Robert L. Thorndike. “Who belongs in the family?” In: *Psychometrika* 18.4 (Dec. 1953). Publisher: Springer, pp. 267–276. DOI: 10.1007/BF02289263. URL: https://link.springer.com/article/10.1007/BF02289263 (cit. on p. 64).

[266] Peter J. Rousseeuw. “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20 (C Nov. 1987). Publisher: North-Holland, pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7 (cit. on p. 64).

[267] T. Caliński and J. Harabasz. “A dendrite method for cluster analysis”. In: *Communications in Statistics - Theory and Methods* 3.1 (1974). Publisher: Marcel Dekker, Inc., pp. 1–27. URL: https://www.tandfonline.com/doi/abs/10.1080/03610927408827101 (cit. on p. 64).

[268] Pavel V Kolesnichenko, Qianhui Zhang, Changxi Zheng, Michael S Fuhrer, and Jeffrey A Davis. "Multidimensional analysis of excitonic spectra of monolayers of tungsten disulphide: toward computer-aided identification of structural and environmental perturbations of 2D materials". In: *Machine Learning: Science and Technology* 2.2 (Mar. 2021). Publisher: IOP Publishing, pp. 025021–025021. DOI: 10.1088/2632-2153/abd87c. URL: https://iopscience.iop.org/article/10.1088/2632-2153/abd87c (cit. on pp. 64, 84).

[269] Adeiza James Onumanyi, Daisy Nkele Molokomme, Sherrin John Isaac, and Adnan M. Abu-Mahfouz. "AutoElbow: An Automatic Elbow Detection Method for Estimating the Number of Clusters in a Dataset". In: *Applied Sciences* 12.15 (Jan. 2022). Number: 15 Publisher: Multidisciplinary Digital Publishing Institute, p. 7515. ISSN: 2076-3417. DOI: 10.3390/app12157515. URL: https://www.mdpi.com/2076-3417/12/15/7515 (cit. on pp. 64, 84, 85).

[270] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. "Distant supervision for relation extraction via Piecewise Convolutional Neural Networks". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 1753–1762. ISBN: 978-1-941643-32-7. DOI: 10.18653/v1/d15-1203. URL: https://aclanthology.org/D15-1203 (cit. on p. 65).

[271] Kai Zhang, Yuan Yao, Ruobing Xie, Xu Han, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. "Open Hierarchical Relation Extraction". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 5682–5693. DOI: 10.18653/v1/2021.naacl-main.452. URL: https://aclanthology.org/2021.naacl-main.452 (cit. on p. 66).

[272] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books". In: Proceedings of the IEEE International Conference on Computer Vision. 2015, pp. 19–27. URL: https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Zhu_Aligning_Books_and_ICCV_2015_paper.html (cit. on p. 68).

[273] Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. "DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2023. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 16207–16221. DOI: 10.18653/v1/2023.acl-long.896. URL: https://aclanthology.org/2023.acl-long.896 (cit. on p. 69).

[274] Taylor Shin, Yasaman Razeghi, Robert L. Logan, Eric Wallace, and Sameer Singh. "AUTOPROMPT: Eliciting knowledge from language models with automatically generated prompts". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Online: Association for Computational Linguistics, Oct. 2020, pp. 4222–4235. ISBN: 978-1-952148-60-6. DOI: 10.18653/v1/2020.emnlp-main.346. URL: https://arxiv.org/abs/2010.15980v2 (cit. on p. 70).

[275] Adi Haviv, Jonathan Berant, and Amir Globerson. "BERTese: Learning to speak to BERT". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2021, pp. 3618–3623. ISBN: 978-1-954085-02-2. DOI: 10.18653/v1/2021.eacl-main.316. URL: https://aclanthology.org/2021.eacl-main.316 (cit. on p. 70).

[276] Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. "CrossNER: Evaluating Cross-Domain Named Entity Recognition". In: *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. Vol. 15. Association for the Advancement of Artificial Intelligence, Dec. 2021, pp. 13452–13460. ISBN: 978-1-71383-597-4. DOI: 10.48550/arxiv.2012.04373. URL: https://arxiv.org/abs/2012.04373v2 (cit. on pp. 77, 79, 89).

[277] Jinyuan Fang, Xiaobin Wang, Zaiqiao Meng, Pengjun Xie, Fei Huang, and Yong Jiang. "MANNER: A Variational Memory-Augmented Model for Cross Domain Few-Shot Named Entity Recognition". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2023. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 4261–4276. DOI: 10.18653/v1/2023.acl-long.234. URL: https://aclanthology.org/2023.acl-long.234 (cit. on pp. 77, 78, 79, 83, 87, 97).

[278] Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. "PromptNER: Prompt Locating and Typing for Named Entity Recognition". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2023. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 12492–12507. DOI: 10.18653/v1/2023.acl-long.698. URL: https://aclanthology.org/2023.acl-long.698 (cit. on pp. 78, 79, 81, 83, 87).

[279] Mozhi Zhang, Hang Yan, Yaqian Zhou, and Xipeng Qiu. *PromptNER: A Prompting Method for Few-shot Named Entity Recognition via k Nearest Neighbor Search*. May 20, 2023. DOI: 10.48550/arXiv.2305.12217. arXiv: 2305.12217[cs]. URL: http://arxiv.org/abs/2305.12217 (cit. on pp. 78, 79).

[280] Tingting Ma, Huiqiang Jiang, Qianhui Wu, Tiejun Zhao, and Chin-Yew Lin. "Decomposed Meta-Learning for Few-Shot Named Entity Recognition". In: *Findings of the Association for Computational Linguistics: ACL 2022*. Findings 2022. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1584–1596. DOI: 10.18653/v1/2022.findings-acl.124. URL: https://aclanthology.org/2022.findings-acl.124 (cit. on pp. 78, 79, 83).

[281] Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca Passonneau, and Rui Zhang. "CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2022. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6338–6353. DOI: 10.18653/v1/2022.acl-long.439. URL: https://aclanthology.org/2022.acl-long.439 (cit. on pp. 78, 86).

[282] Yucheng Huang, Kai He, Yige Wang, Xianli Zhang, Tieliang Gong, Rui Mao, and Chen Li. "COPNER: Contrastive Learning with Prompt Guiding for Few-shot Named Entity Recognition". In: *Proceedings of the 29th International Conference on Computational Linguistics*. COLING 2022. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 2515–2527. URL: https://aclanthology.org/2022.coling-1.222 (cit. on pp. 78, 86).

[283] Guanting Dong, Zechen Wang, Jinxu Zhao, Gang Zhao, Daichi Guo, Dayuan Fu, Tingfeng Hui, Chen Zeng, Keqing He, Xuefeng Li, Liwen Wang, Xinyue Cui, and Weiran Xu. "A Multi-Task Semantic Decomposition Framework with Task-specific Pre-training for Few-Shot NER". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. CIKM '23. New York, NY, USA: Association for Computing Machinery, Oct. 21, 2023, pp. 430–440. DOI: 10.1145/3583780.3614766. URL: https://doi.org/10.1145/3583780.3614766 (cit. on pp. 78, 79, 83).

[284] PENG Fuchun. "Accurate information extraction from research papers using conditional random fields". In: *Proc. of HLT/NAACL, 2004*. 2004 (cit. on p. 79).

[285] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. "Named Entity Recognition as Structured Span Prediction". In: *Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)*. UM-IoS 2022. Ed. by Wenjuan Han, Zilong Zheng, Zhouhan Lin, Lifeng Jin, Yikang Shen, Yoon Kim, and Kewei Tu. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 1–10. DOI: 10.18653/v1/2022.umios-1.1. URL: https://aclanthology.org/2022.umios-1.1 (cit. on p. 79).

[286] Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. "OpenPrompt: An Open-source Framework for Prompt-learning". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Ed. by Valerio Basile, Zornitsa Kozareva, and Sanja Stajner. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 105–113. DOI: 10.18653/v1/2022.acl-demo.10. URL: https://aclanthology.org/2022.acl-demo.10 (cit. on p. 79).

[287] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. Proceedings of Machine Learning Research, July 17, 2017, pp. 1126–1135. URL: https://proceedings.mlr.press/v70/finn17a.html (cit. on p. 79).

[288] Aniruddha Mahapatra, Sharmila Reddy Nangi, Aparna Garimella, and Anandhavelu Natarajan. “Entity Extraction in Low Resource Domains with Selective Pre-training of Large Language Models”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 942–951. URL: https://aclanthology.org/2022.emnlp-main.61 (cit. on pp. 79, 80).

[289] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. “Self-Consistency Improves Chain of Thought Reasoning in Language Models”. In: *Proceedings of the 11th International Conference on Learning Representations*. 11th International Conference on Learning Representations. Kigali, Rwanda, Mar. 7, 2023. DOI: 10.48550/arXiv.2203.11171. URL: http://arxiv.org/abs/2203.11171 (cit. on p. 80).

[290] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. “Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023”. In: *URL https://lmsys. org/blog/2023-03-30-vicuna* 3.5 (2023) (cit. on p. 80).

[291] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. *Mistral 7B*. Oct. 10, 2023. DOI: 10.48550/arXiv.2310.06825. arXiv: 2310.06825[cs]. URL: http://arxiv.org/abs/2310.06825 (cit. on p. 80).

[292] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. *Code Llama: Open Foundation Models for Code*. Jan. 31, 2024. DOI: 10.48550/arXiv.2308.12950. arXiv: 2308.12950[cs]. URL: http://arxiv.org/abs/2308.12950 (cit. on p. 80).

[293] Pengcheng He, Jianfeng Gao, and Weizhu Chen. “DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing”. In: *Proceedings of the Eleventh International Conference on Learning Representations*. The Eleventh International Conference on Learning Representations. Kigali, Rwanda, 2023. URL: https://openreview.net/forum?id=sE7-XhLxHA (cit. on pp. 80, 87, 90, 104).

[294] Alessandro Cucchiarelli and Paola Velardi. “Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence”. In: *Computational Linguistics* 27.1 (Mar. 1, 2001), pp. 123–131. ISSN: 0891-2017. DOI: 10.1162/089120101300346822. URL: https://doi.org/10.1162/089120101300346822 (cit. on p. 81).

[295] David Nadeau, Peter Turney, and Stan Matwin. “Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity”. In: *Advances in Artificial Intelligence*. Ed. by Luc Lamontagne and Mario Marchand. Lecture Notes in Computer

Science. Berlin, Heidelberg: Springer, 2006, pp. 266–277. ISBN: 978-3-540-34630-2. DOI: 10.1007/11766247_23 (cit. on p. 81).

[296] Zihan Liu, Genta Indra Winata, and Pascale Fung. "Zero-Resource Cross-Domain Named Entity Recognition". In: *Proceedings of the 5th Workshop on Representation Learning for NLP*. RepL4NLP 2020. Ed. by Spandana Gella, Johannes Welbl, Marek Rei, Fabio Petroni, Patrick Lewis, Emma Strubell, Minjoon Seo, and Hannaneh Hajishirzi. Online: Association for Computational Linguistics, July 2020, pp. 1–6. DOI: 10.18653/v1/2020.repl4nlp-1.1. URL: https://aclanthology.org/2020.repl4nlp-1.1 (cit. on p. 81).

[297] Ying Luo, Hai Zhao, and Junlang Zhan. "Named Entity Recognition Only from Word Embeddings". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. EMNLP 2020. Online: Association for Computational Linguistics, Nov. 2020, pp. 8995–9005. DOI: 10.18653/v1/2020.emnlp-main.723. URL: https://aclanthology.org/2020.emnlp-main.723 (cit. on pp. 81, 83, 87).

[298] Richard Duda, Peter Hart, and David Stork. *Pattern classification*. Wiley Hoboken, 2000. ISBN: 0-471-05669-3 (cit. on p. 84).

[299] Gideon Schwarz. "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6.2 (1978). Publisher: Institute of Mathematical Statistics, pp. 461–464. ISSN: 0090-5364. URL: https://www.jstor.org/stable/2958889 (cit. on pp. 85, 111).

[300] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. "Large Scale Online Learning of Image Similarity Through Ranking". In: *The Journal of Machine Learning Research* 11 (Mar. 1, 2010), pp. 1109–1135. ISSN: 1532-4435. URL: https://www.jmlr.org/papers/volume11/chechik10a/chechik10a.pdf (cit. on pp. 86, 111).

[301] Changyou Chen, Jianyi Zhang, Yi Xu, Liqun Chen, Jiali Duan, Yiran Chen, Son Tran, Belinda Zeng, and Trishul Chilimbi. "Why do We Need Large Batchsizes in Contrastive Learning? A Gradient-Bias Perspective". In: *Advances in Neural Information Processing Systems* 35 (Dec. 6, 2022), pp. 33860–33875 (cit. on p. 86).

[302] Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. "Few-shot Slot Tagging with Collapsed Dependency Transfer and Label-enhanced Task-adaptive Projection Network". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 1381–1393. DOI: 10.18653/v1/2020.acl-main.128. URL: https://aclanthology.org/2020.acl-main.128 (cit. on p. 87).

[303] Jingjing Liu, Panupong Pasupat, Scott Cyphers, and Jim Glass. "Asgard: A portable architecture for multilingual dialogue systems". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. ISSN: 2379-190X. May 2013, pp. 8386–8390. DOI: 10.1109/ICASSP.2013.6639301. URL: https://ieeexplore.ieee.org/abstract/document/6639301 (cit. on p. 89).

[304] Aman Kumar and Binil Starly. ""FabNER": information extraction from manufacturing process science domain literature using named entity recognition". In: *Journal of Intelligent Manufacturing* 33.8 (Dec. 1, 2022), pp. 2393–2407. ISSN: 1572-8145. DOI: 10.1007/s10845-021-01807-x. URL: https://doi.org/10.1007/s10845-021-01807-x (cit. on p. 89).

[305] Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. "GENIA corpus—a semantically annotated corpus for bio-textmining". In: *Bioinformatics* 19 (suppl_1 2003). ISBN: 1367-4811 Publisher: Oxford University Press, pp. i180–i182 (cit. on p. 89).

[306] Amber Stubbs and Özlem Uzuner. "Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus". In: *Journal of Biomedical Informatics* 58 Suppl (Suppl Dec. 2015), S20–S29. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2015.07.020 (cit. on p. 89).

[307] Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. "GENTLE: A Genre-Diverse Multilayer Challenge Set for English NLP and Linguistic Evaluation". In: *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*. LAW 2023. Ed. by Jakob Prange and Annemarie Friedrich. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 166–178. DOI: 10.18653/v1/2023.law-1.17. URL: https://aclanthology.org/2023.law-1.17 (cit. on p. 89).

[308] Amir Zeldes. "The GUM corpus: creating multilayer resources in the classroom". In: *Language Resources and Evaluation* 51.3 (Sept. 1, 2017), pp. 581–612. ISSN: 1574-0218. DOI: 10.1007/s10579-016-9343-x. URL: https://doi.org/10.1007/s10579-016-9343-x (cit. on p. 89).

[309] Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. "Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition". In: *Proceedings of the 3rd Workshop on Noisy User-generated Text*. WNUT 2017. Ed. by Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 140–147. DOI: 10.18653/v1/W17-4418. URL: https://aclanthology.org/W17-4418 (cit. on p. 89).

[310] Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. *Train Once, Test Anywhere: Zero-Shot Learning for Text Classification*. Dec. 23, 2017. DOI: 10.48550/arXiv.1712.05972. arXiv: 1712.05972[cs]. URL: http://arxiv.org/abs/1712.05972 (cit. on p. 90).

[311] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. "DeBERTa: Decoding-enhanced BERT with Disentangled Attention". In: *Proceedings of the Ninth International Conference on Learning Representations*. The Ninth International Conference on Learning Representations. Online, Jan. 12, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD (cit. on pp. 90, 104).

[312] Diederik Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization." In: *Proceedings of the 3rd International Conference on Learning Representations, Conference Track*. 3rd International Conference on Learning Representations. San Diego, CA, USA, 2015. URL: http://arxiv.org/abs/1412.6980 (cit. on pp. 90, 91).

[313] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008) (cit. on p. 99).

[314] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators". In: *Proceedings of the Seventh International Conference on Learning Representations*. The Seventh International Conference on Learning Representations. New Orleans, Louisiana, United States, Sept. 25, 2019. URL: https://openreview.net/forum?id=r1xMH1BtvB (cit. on p. 104).

[315] Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. "ReFinED: An Efficient Zero-shot-capable Approach to End-to-End Entity Linking". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*. NAACL-HLT 2022. Ed. by Anastassia Loukina, Rashmi Gangadharaiah, and Bonan Min. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, July 2022, pp. 209–220. DOI: 10.18653/v1/2022.naacl-industry.24. URL: https://aclanthology.org/2022.naacl-industry.24 (cit. on p. 112).

[316] Karl Pearson and John Blakeman. *On the theory of contingency and its relation to association and normal correlation*. Vol. XIII. Mathematical contributions to the theory of evolution. 1904 (cit. on p. 116).

[317] Roy Jonker and Ton Volgenant. "A shortest augmenting path algorithm for dense and sparse linear assignment problems". In: *DGOR/NSOR: Papers of the 16th annual meeting of DGOR in cooperation with NSOR/vorträge der 16. Jahrestagung der DGOR zusammen mit der NSOR*. Springer, 1988, pp. 622–622 (cit. on p. 116).

[318] David F. Crouse. "On implementing 2D rectangular assignment algorithms". In: *IEEE Transactions on Aerospace and Electronic Systems* 52.4 (Aug. 2016). Conference Name: IEEE Transactions on Aerospace and Electronic Systems, pp. 1679–1696. ISSN: 1557-9603. DOI: 10.1109/TAES.2016.140952. URL: https://ieeexplore.ieee.org/document/7738348 (cit. on p. 116).

[319] E. Cuthill and J. McKee. "Reducing the bandwidth of sparse symmetric matrices". In: *Proceedings of the 1969 24th national conference*. ACM '69. New York, NY, USA: Association for Computing Machinery, 1969, pp. 157–172. ISBN: 978-1-4503-7493-4. DOI: 10.1145/800195.805928. URL: https://dl.acm.org/doi/10.1145/800195.805928 (cit. on p. 118).

[320] Alan George and Joseph W Liu. *Computer solution of large sparse positive definite*. Prentice Hall Professional Technical Reference, 1981 (cit. on p. 118).

FOLIO ADMINISTRATIF

THESE DE L'INSA LYON, MEMBRE DE L'UNIVERSITE DE LYON

NOM : Genest DATE de SOUTENANCE : 09/12/2024

Prénoms : Pierre-Yves

TITRE : Unsupervised Open-World Information Extraction From Unstructured and Domain-Specific Document Collections

NATURE : Doctorat Numéro d'ordre : 2024ISAL0111

Ecole doctorale : InfoMaths

Spécialité : Informatique

RESUME : The exponential growth in data generation has rendered the effective analysis of unstructured textual document collections a critical challenge. This PhD thesis aims to address this challenge by focusing on Information Extraction (IE), which encompasses four essential tasks: Named Entity Recognition (NER), Coreference Resolution (CR), Entity Linking (EL), and Relation Extraction (RE). These tasks collectively enable extracting and structuring knowledge from unformatted documents, facilitating its integration into structured databases for further analytical processes.

Our contributions start with creating Linked-DocRED, the first large-scale, diverse, and manually annotated dataset for document-level IE. This dataset enriches the existing DocRED dataset with high-quality entity linking labels. Additionally, we propose a novel set of metrics for evaluating end-to-end IE models. The evaluation of baseline models on Linked-DocRED highlights the complexities and challenges inherent to document-level IE: cascading errors, long context handling, and information scarcity.

We then introduce PromptORE, an unsupervised and open-world RE model. Adapting the prompt-tuning paradigm, PromptORE achieves relation embedding and clustering without requiring fine-tuning or hyperparameter tuning (a major weakness of previous baselines) and significantly outperforms state-of-the-art models. This method demonstrates the feasibility of extracting semantically coherent relation types in an open-world context.

Further extending our prompt-based approach, we develop CITRUN for unsupervised and open-world NER. By employing contrastive learning with off-domain labeled data, CITRUN improves entity type embeddings, surpassing LLM-based unsupervised NERs, and achieving competitive performance against zero-shot models that are more supervised.

These advancements facilitate meaningful knowledge extraction from unstructured documents, addressing practical, real-world constraints and enhancing the applicability of IE models in industrial contexts.

MOTS-CLÉS : information extraction, open named entity recognition, unsupervised named entity recognition, open relation extraction, unsupervised relation extraction, natural language processing.

Laboratoire de recherche : Laboratoire d'Informatique en Image et Systèmes d'information (LIRIS), UMR 5205

Directeur de thèse : Előd Egyed-Zsigmond

Présidente de jury : Josiane Mothe

Composition du jury : Josiane Mothe (présidente), Gabriele Gianini (rapporteur), Michael Granitzer (rapporteur), Sylvie Calabretto (examinatrice), Előd Egyed-Zsigmond (directeur de thèse), Pierre-Edouard Portier (invité, co-encadrant), Martino Lovisetto (invité)