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

Pierre-Yves Genest, Pierre-Edouard Portier, Elöd Egyed-Zsigmond, Martino Lovisetto

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

Pierre-Yves Genest[1, 2], Pierre-Edouard Portier[2], Elődo Egyed-Zsigmond[2], and Martino Lovisetto[1]

[1]Alteca, 88 Boulevard des Belges, 69006 Lyon, France
[2]Univ Lyon, INSA Lyon, CNRS, UCBL, LIRIS, UMR5205, 20 Avenue Einstein, 69621 Villeurbanne, France
{*pygenest,mlovisetto*}*@alteca.fr* {*pierre-edouard.portier,elod.egyed-zsigmond*}*@insa-lyon.fr*

February 1, 2023

Abstract

Information Extraction (IE) pipelines aim to extract meaningful entities and relations from documents and structure them into a knowledge graph that can then be used in downstream applications. Training and evaluating such pipelines requires a dataset annotated with entities, coreferences, relations, and entity-linking. However, existing datasets either lack entity-linking labels, are too small, not diverse enough, or automatically annotated (that is, without a strong guarantee of the correction of annotations). Therefore, we propose Linked-DocRED, to the best of our knowledge, the first manually-annotated, large-scale, document-level IE dataset. 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 disambiguation candidates. We also propose a complete framework of metrics to benchmark end-to-end IE pipelines, and we define an entity-centric metric to evaluate entity-linking. The evaluation of a baseline shows promising results while highlighting the challenges of an end-to-end IE pipeline. 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].

Keywords— information extraction; document-level relation extraction; entity-linking; dataset

1 Introduction

Information Extraction (IE) aims to extract the meaningful information from documents, that is, entities and relations between these entities, to build or complement a Knowledge Graph (KG). The resulting knowledge graph can then be used for multiple downstream tasks such as recommender systems [14], logical reasoning [4], or question answering [17]. Similarly to [37, 28], we define IE as a four-step process with:

1. Named Entity Recognition (NER),
2. Coreference Resolution (Coref),
3. Relation Extraction (RE),
4. Entity-Linking (EL).

Information extraction can be seen as a supervised task [37, 28, 32], a weakly-supervised task [10], or an unsupervised task [13, 2], the most common setting being supervised information extraction. Several datasets have been proposed to train and evaluate such pipelines. The most recent ones [36, 37, 23] focus on document-level information extraction, a more realistic, albeit more challenging scenario than sentence-level IE.

However, none of these datasets is entirely satisfactory for the end-to-end evaluation of IE pipelines, covering the four steps: NER, RE, Coref, and EL. On the one hand, most datasets focus on NER, Coref, and RE, ignoring the last entity-linking step [36, 5]. Nonetheless, entity-linking is one of the most important steps, if not the most important, as it transforms ambiguous extracted triples into structured and disambiguated nodes and relations. The question of ambiguity in natural language is essential: 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.

On the other hand, datasets that provide entity-linking annotations are either too small, not diverse enough, too simple (e.g., using sentences and not documents), or automatically annotated [23, 37, 10, 12].

Therefore, we propose **Linked-DocRED**, to the best of our knowledge, the first large-scale, manually labeled, document-level IE dataset that provides annotations for entities, coreferences, relations, and entity-linking. Linked-DocRED aims to correct the shortcomings of existing datasets

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

and to define a reproducible and more complete benchmark for the training and evaluation of end-to-end IE pipelines.

Instead of creating a dataset from scratch, we enhance the widely-used DocRED dataset [36] (already labeled with entities, coreferences, and relations) by annotating each entity with entity-linking. Since DocRED documents are taken from Wikipedia articles, we propose to use 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 [5]).

We also continue the work of Zaporojets et al. [37] by establishing a clear and coherent set of entity-centric metrics to evaluate the performance of an IE pipeline. In particular, we define an entity-centric metric to assess entity-linking. The evaluation of a baseline method based on recent approaches shows encouraging results but also demonstrates that this task is still a difficult challenge, in particular, because of cascading errors during successive steps of an IE pipeline. We hope that Linked-DocRED can facilitate the discovery of more performant IE pipelines.

Let us 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 [37].
- We propose a new entity-linking method based on the alignment between DocRED documents and Wikipedia articles, providing high-quality labeling, a method that can be applied to disambiguate other Wikipedia-based datasets.
- We define a novel entity-centric metric to assess entity-linking in order to provide a complete set of metrics to evaluate an IE pipeline.
- We adapt state-of-the-art approaches to provide a simple and reproducible baseline covering the four steps of an IE pipeline, namely NER, Coref, RE, and EL. The experimental results are promising, with, however, a large margin of progress, in particular for entity-linking, which, at the end of the pipeline, is subject to the effects of cascading errors.

2 Related Work

As we have said in the introduction, we define information extraction as a four-step process with [28, 37]:

1. Named Entity Recognition – extracting and typing the surface forms of entities in a piece of text,
2. Coreference Resolution – identifying the surface forms that refer to the same entity in a piece of text,
3. Relation Extraction – extracting and typing the relations occurring between the extracted entities in a piece of text,
4. Entity-Linking or Entity Disambiguation – identifying, for an extracted entity, the corresponding resource in a predetermined knowledge graph.

Recent papers often consider the first three tasks [40, 38, 16, 20, 33, 31], setting entity-linking aside. To the best of our knowledge, only a handful of papers [32, 28, 9] are exploring the end-to-end pipeline. In our opinion, 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 pipelines, numerous datasets have been proposed, covering a large spectrum of settings and applications:

- Some focus on general domain information (e.g., T-REx [10], DocRED [36], or HacRED [5]), other on very specific domains (scientific literature for SciERC [20], biomedicine for FewRel 2.0 [12], or BC5CDR [19]).
- Some are manually annotated (e.g., DocRED [36], FewRel [15] or HacRED [5]), others automatically generated such as T-REx [10] or NYT-10 [29].
- Some focus on sentences (e.g., FewRel [15, 12], or NYT-10 [29]) others on documents (DocRED [36], KnowledgeNet [23], or DWIE [37]).

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

FewRel [15, 12] It is large-scale, diverse (it contains many different relation types), and annotated for the four tasks, but it 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 [10] Of the seven datasets, it is by far the largest, with around 4.6 million documents. It is not usable in our scenario, though, as the dataset was automatically labeled, which means there is no strong guarantee of the quality of annotations. Nevertheless, it provides a huge source of distant-supervision, which can be beneficial during training (even though it is a lower-quality annotation).

KnowledgeNet [23] and BC5CDR [19] They contain documents and are annotated for the four tasks. Similarly to FewRel, BC5CDR has no new knowledge (all entities are already present in the knowledge base). This default is absent of KnowledgeNet, with an appreciable presence of new knowledge. However, BC5CDR and KnowledgeNet are too small (see Table 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.

DWIE [37] Similar to KnowledgeNet and BC5CDR, DWIE contains documents labeled for the four tasks. It is more diverse and bigger, though, but still a lot smaller in

Table 1: Quantitative comparison of Linked-DocRED and widely-used IE datasets. *# Entities*: number of entities in the documents, ignoring coreferences; *# Coref.*: number of coreferences; *Entity-Linking # New*: number of entities that do not exist in the reference knowledge graph; *Relation # Inst.*: number of relations between entities, ignoring coreferences.

Dataset	Size		Entities		# Coref.	Entity-Linking		Relations	
	# Docs	# Tokens	# Entities	# Types		# Linked	# New	# Inst.	# Types
FewRel [15, 12]	-	1 397k	112k	-	2k	112k	0	56k	80
T-REx [10]	4 650.0k	446 053k	69 962k	-	17 617k	69 962k	0	208 774k	642
KnowledgeNet [23]	4.0k	734k	11k	-	7k	9k	1.9k	13k	15
BC5CDR [19]	1.5k	343k	10k	2	19k	10k	0	48k	1
DWIE [37]	0.8k	501k	23k	311	20k	13k	10.0k	22k	65
HacRED [5]	9.2k	1 141k	99k	9	19k	-	-	68k	26
DocRED [36]	5.1k	1 001k	99k	6	34k	-	-	50k	96
Linked-DocRED	5.1k	1 001k	95k	6	38k	63k	6.4k	50k	96

terms of documents, entities, and relations compared to HacRED and DocRED. In any case, the analysis of the dataset's files suggests that entity-linking was automatically labeled (multiple candidates with eighteen-digit precision probabilities). As a result, it is not satisfactory for our purpose.

DocRED [36] and HacRED [5] They contain around two to five times more documents and annotations than the other manually annotated datasets, which makes them more suitable to train and evaluate IE pipelines. Unfortunately, they are not annotated with entity-linking.

Although several datasets have been proposed to evaluate IE pipelines, none is entirely satisfactory. Indeed, FewRel [15, 12] lacks documents and novel entities; T-REx [10] lacks manual annotations and novel entities; KnowledgeNet [23] and BC5CDR [19] are too small and not diverse enough; DWIE has automatic entity-linking annotations [37]; and HacRED [5], and DocRED [36] lack annotation 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 pipelines.

3 Dataset Generation

In this section, we describe the process we used to create Linked-DocRED. First, creating an IE dataset from scratch is a very expensive enterprise as it requires annotating documents for entities, coreferences, relations, and entity-linking. In particular, entity-linking is very time-consuming due to 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 [36], is almost adequate to train and evaluate an IE pipeline, 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 entity-linker (for instance DBPedia Spotlight [22]), as they would introduce biases. Indeed, entity-linkers are imperfect (in fact, even human annotation is imperfect) and have advantages and drawbacks. So, if an IE pipeline uses the same entity-linker for its predictions, it will reproduce the same behavior and obtain overstated (biased) results. The only valid choice for us is to rely on manual annotations to limit the introduction of bias in Linked-DocRED.

Our entity-linking aims to link every entity of DocRED to a resource in Wikipedia. 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*[2] in Figure 5). We will also provide the Wikidata identifier associated with the Wikipedia resource. Providing these two identifiers is beneficial: 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, that is, the first paragraphs of a Wikipedia article. If we take the instance presented in Figure 5, the document corresponds to the Wikipedia abstract of Luke Skywalker[3]. The hyperlinks in the Wikipedia article 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, to be more precise, a form of manual entity-linking because Wikipedia contributors manually edit these hyperlinks. Besides, we note that there is a direct mapping (same sentence, same position) between a lot of DocRED entities in the document and hyperlinks in the corresponding Wikipedia article (e.g., *Star Wars*, *George Lucas*, *Mark Hamill*, *Padmé Amidala*, or *Galactic Empire* in Figure 5). Using these hyperlinks 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 1. The main step is mapping entities with Wikipedia hyperlinks, which is the second module of Figure 1 (Hyperlinks Alignment). It is not sufficient to fully disambiguate our dataset, which explains the three steps that follow it. In the next parts, we will describe each constituent in the disambiguation process for a DocRED

[2] This entity is not in Wikipedia at the time we write this article.

[3] Available at `https://en.wikipedia.org/wiki/Luke_Skywalker`.

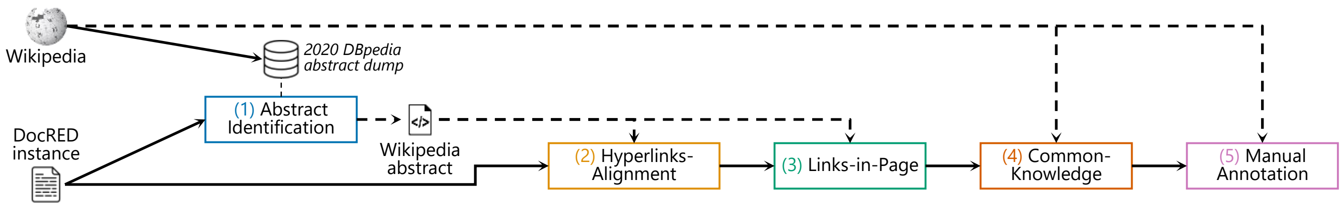

Figure 1: Architecture of the semi-automatic entity-linking process implemented to disambiguate Linked-DocRED.

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, 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.

3.1 Wikipedia Abstract Identification

To access the hyperlinks and map them to our entities, we first need to get the Wikipedia article associated with our DocRED document (the first step in Figure 1). Although DocRED does not contain the URL of the source Wikipedia page, we have access to the article title and, obviously, the abstract text. A possible solution is to do a full-text search on the title or abstract to find the most similar Wikipedia article.

A second 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. To mitigate this issue, we downloaded the 2020-01 DBPedia abstracts dump[4], which is the oldest available this day. We have also tested with Wikipedia dumps, but we found them of lower quality (some abstracts were truncated, and others contained abnormal characters). From the DBpedia dump, 5.6M of Wikipedia abstracts were indexed in ElasticSearch[5].

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 bag-of-words and the BM25 metric [30] to perform its full-text search. This setup is very efficient and fast in returning good Wikipedia candidates, but it does not consider the ordering of the words in the Wikipedia article. To have the best confidence possible, we propose to rank the candidates using a similarity metric based on the Levenshtein distance [18] (which measures the number of modifications to make to transform the first string into the second):

$$\mathrm{sim}_{\mathrm{text}}(t_1, t_2) = 1 - \frac{\mathrm{d}_{Levenshtein}(t_1, t_2)}{\max(\mathrm{len}(t_1), \mathrm{len}(t_2))}, \quad (1)$$

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

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

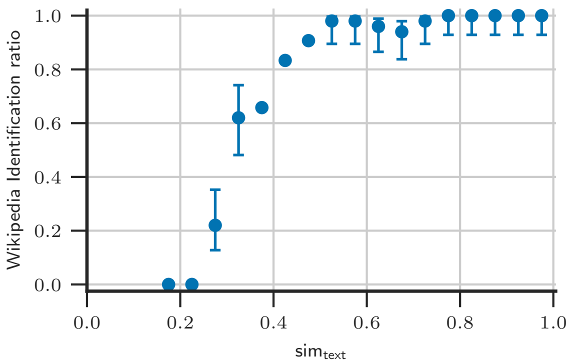

Figure 2: Results of the manual annotation to determine the threshold of $\mathrm{sim}_{\mathrm{text}}$ to maximize the correct Wikipedia article identification. We show the confidence interval with $\alpha = 0.05$ (no confidence interval if all instances of the bin have been annotated).

where t_1 and t_2 are the two strings to be compared, *len* computes the length of a string, and $d_{Levenshtein}$ is the Levenshtein distance.

Although this similarity metric ranks precisely the candidates (logically, the DocRED document and the correct Wikipedia candidate are the closest in terms of editing distance), it cannot determine whether the first Wikipedia candidate is the right article. Indeed, our DBPedia dump is incomplete[6]: it does not contain every Wikipedia abstract, which means that some DocRED documents cannot be found. To filter those instances, we propose to determine a threshold with our similarity metric. We select a sample of 1 000 DocRED documents with their first Wikipedia candidate, stratified with $\mathrm{sim}_{\mathrm{text}}$ (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 2. For each bin, we also compute the confidence interval for the proportion with $\alpha = 0.05$, using the Wilson approximation [35], due to the low number of samples per bin and the proportion being close to 0 or 1. 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.

In Figure 2, for $\mathrm{sim}_{\mathrm{text}} > 0.5$, the proportion of correctly identified Wikipedia articles is close to 1 (above 0.95). Therefore, we propose to set our threshold at $\mathrm{sim}_{\mathrm{text}} > 0.5$ and check DocRED documents with $\mathrm{sim}_{\mathrm{text}} \leq 0.5$ manually. Using this threshold, we automatically identify the Wikipedia

[6] At the time we write this article, there are around 6.6M Wikipedia articles compared to the 5.6M in the DBpedia dump.

article for 4 694 documents (93%). We manually determine the Wikipedia abstract for the remaining 357 documents and could not find the Wikipedia article for 23 instances (we think these articles have been completely removed from Wikipedia).

3.2 Hyperlinks Alignment

In this module (second step in Figure 1), we implement the mapping between the DocRED document's entities and hyperlinks in the Wikipedia article we have found previously. We want to find direct intersections (same sentence and position) between entities in our DocRED instance and hyperlinks in the Wikipedia article. To do that, we need to align precisely our DocRED text with the Wikipedia abstract. The problem is that there are minor differences between the two texts (due to the preprocessing applied on DocRED instances that removes Cyrillic, Arabic, and Asiatic characters; or some parts of the abstract), which make this step nontrivial.

To overcome this difficulty, we propose to use the Needleman-Wunsch algorithm [25], which was initially proposed to optimally align two nearly-identical DNA sequences, allowing insertions, deletions, and substitutions of nucleotides. This algorithm is easily generalizable to string alignment by replacing the notion of nucleotides with characters. It allows us to produce a translation table to convert a character position in the DocRED instance to a position in the Wikipedia article. Once we have this translation table, it is simple to compute intersections between surface forms of entities as annotated in DocRED and Wikipedia hyperlinks and thus generate candidate entity-linkings.

We have, however, no warranty on the quality of the proposed candidates. Intuitively, if the intersection is exact, the entity-linking should be accurate, but it becomes more difficult with a partial intersection (e.g., *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 simple measure could be 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 3.1. We first compute $\mathrm{sim}_{\mathrm{text}}$ (see Eq. 1) between the DocRED entity text and the matched Wikipedia hyperlink, which allows us to quantify the quality of the intersection. We then select a sample of 1 000 entities and their matched hyperlinks, stratified on $\mathrm{sim}_{\mathrm{text}}$ (with 20 bins of 0.05), and manually determine whether the entity-linking is correct. For each bin, we also compute a confidence interval for a proportion with $\alpha = 0.05$. The results are shown in Figure 3.

We can see three regimes:

- $\mathrm{sim}_{\mathrm{text}} < 0.35$ – few entities are correctly disambiguated, which is logical given that the entity and the hyperlink are dissimilar,
- $\mathrm{sim}_{\mathrm{text}} \in [0.35, 0.75]$ – entities and hyperlinks are relatively similar, but the probability of wrong entity-linking is still high,
- $\mathrm{sim}_{\mathrm{text}} > 0.75$ – the proportion is close to 1 (0.984): of the 250 annotated pairs, only four are wrongly linked.

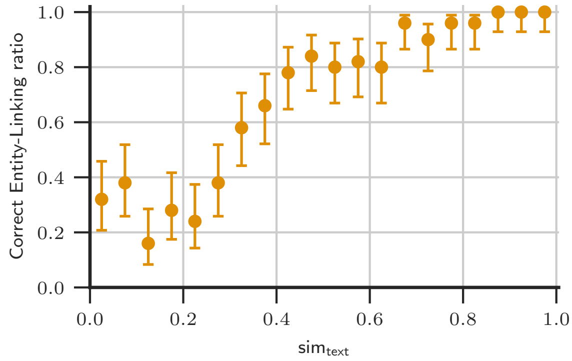

Figure 3: Results of the manual annotation to evaluate the disambiguation quality depending on the $\mathrm{sim}_{\mathrm{text}}$ between the entity and the hyperlink. We show the confidence interval with $\alpha = 0.05$.

Considering this Figure 3, we decide to keep only entity-linking candidates with the highest entity-linking proportion, that is, with $\mathrm{sim}_{\mathrm{text}} > 0.75$. By doing so, we disambiguate 40 826 entities of DocRED (43.3%) as shown in Figure 4.

This module provides annotations with high confidence, as we are 1. very strict with intersections and textual similarity, and 2. relying on manual annotations of Wikipedia contributors.

3.3 Links in Page

In a Wikipedia article, the first mention of an entity is associated with a hyperlink, 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 there is a hyperlink on the Wikipedia page with the same surface form as the entity we are disambiguating (the third step in Figure 1). Using this approach, we disambiguate 6 741 additional entities (7.1%).

This approach is of lower quality compared to *Hyperlinks Alignment*. However, we are strict on selecting hyperlinks (exact match between the hyperlink and the surface form of the entity).

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 hyperlinks, as it is supposed that everyone knows them already.

To add this notion of common knowledge (the fourth step in Figure 1), we select the entities mentioned at least three times in the dataset and manually annotate them. We take particular care to detect entities with ambiguities, for instance, *French* can refer to France, the French language, or the French people; or *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 the *Links in Page* module, as the two processes are similar.

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[7].

The annotation is done document by document. The annotators must label every remaining entity (three entities 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 a famous web search engine using the surface form and filtering to keep only Wikipedia results. 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[8]: 361 entities were wrongly typed, 148 mentions needed to be corrected (the entity's 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 entity-linkings were similar. On this sample, we compute the Cohen's kappa coefficient [7], and obtain

$$\kappa_{entity-linking} = 0.679.$$

This $\kappa_{entity-linking}$ score shows a strong inter-annotator agreement, especially considering the diversity of Wikipedia resources (more than 6.6M 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 $\kappa_{entity-linking} = 0.816$, which indicates a very strong agreement between annotators.

Overall, this inter-annotator agreement analysis exhibits the high quality of the annotation. 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."

[7]Available at `https://labelstud.io/`.

[8]This error identification step is not exhaustive.

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 were refused by the annotators, as they are candidates that an entity-linker can easily predict, and we are sure that these candidates are wrong.

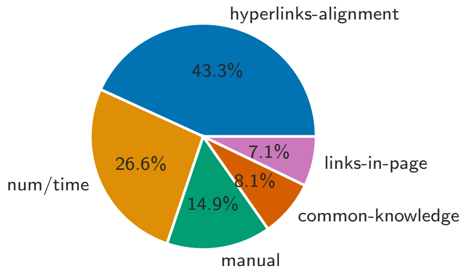

Figure 4: Modules used to disambiguate the 94 547 entities of Linked-DocRED (see section 3 for the details).

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 4.

To find the Wikidata id for each disambiguated entity, we use the metadata of the Wikipedia resource (the property `wikibase_item`).

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 1.

The instance 2774 of the `train` split is shown in Figure 5. 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 / Wikidata (related to the node `#DocRED-6032#`).

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. [36].

Entities & Coreferences Entities are automatically extracted and typed using spaCy[9]. 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

[9]Available at `https://spacy.io/`.

(1) 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.

(2)

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

(3)

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

(4)

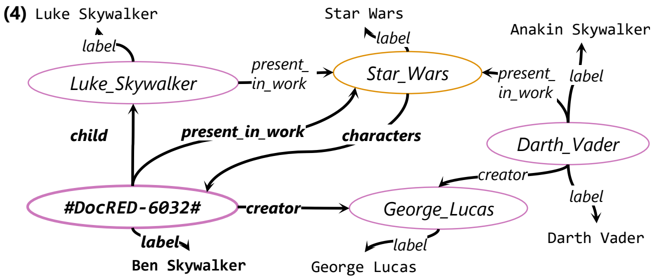

Figure 5: Example instance of Linked-DocRED. From top to bottom: (1) Text of a document with highlighted entities, (2) Two examples of extracted entities with their Wikipedia and Wikidata resources, Ben Skywalker has no corresponding resource in Wikipedia, (3) Two examples of relations, (4) Small part of the knowledge graph built from the entities and relations of the document.

[11]. 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.

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 (e.g., Wikidata), 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. [36]). 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 5).

As we can see in Table 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 2774 of the `train` split (see Figure 5), *Darth Vader* and *Anakin Skywalker* were not identified as coreferences.

4.2 Entity-Linking

The entity-linking annotation process is described in Section 3. To sum up, we rely on human annotations elicited by a semi-automatic process, as shown in Figure 1: (1–3) we map entities with Wikipedia hyperlinks 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 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 that are 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 5.
  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 5.
- `wikidata_resource`: the identifier of the Wikidata entity, for instance `Q12206942` for entity 12 in Figure 5.
- `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 4).
- `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:

- A *(very high confidence)* – if the entity is linked using hyperlinks alignment, manual annotation, or is ignored (NUM and TIME),
- B *(high confidence)* – if the entity is linked using links in page or common knowledge,
- C *(medium confidence)* – if the annotator indicates 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

Table 2: Proportion of Linked-DocRED entities associated with each confidence indicator and estimation of the correct entity-linking probability (the confidence interval with $\alpha = 0.05$ is also shown).

Confidence Indicator	A	B	C
Proportion in Linked-DocRED	78.0%	15.2%	6.8%
Correct entity-linking probability	0.979 ± 0.009	0.950 ± 0.014	-

us to estimate the probability of correct entity-linking and we also compute a 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 2.

We see that 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.

5 Experiments

5.1 Baseline

As we have seen in section 2, an end-to-end IE pipeline can be seen as a four-module process with 1. Named Entity Recognition, 2. Coreference Resolution, 3. Relation Extraction, and 4. Entity-Linking. Our objective for this baseline is to provide a simple IE pipeline with comparable results to current state-of-the-art approaches. Recent papers that use DocRED as a benchmark focus on document-level RE [38, 34, 16, 42], ignoring NER and Coreference Resolution.

Named Entity Recognition We propose to use the simple yet effective span-based NER proposed by Zhong and Chen [41, 33, 21] (PURE). This model relies on BERT [8], 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 [3], which can encode documents up to 4 096 tokens, with only a marginal decrease in performance compared to BERT.

Coreference Resolution We propose to implement a well-used model, NeuralCoref[10]. 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 [26], 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. [28], we propose to use ATLOP [42] to extract relations. Contrary to concurrent approaches (e.g., [38, 39, 37, 40, 6]), who often represent the knowledge explicitly as a graph, which can be processed with Graph Neural Networks (GNN) for inference; Zhou et al. [42] 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-Coref-RE ensemble *PNA* (for PURE [41], NeuralCoref, and ATLOP [42]). This pipeline is trained using the hyperparameter values proposed by the authors of PURE [41], NeuralCoref, and ATLOP [42].

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 pipelines ([28, 37, 32] use very basic approaches).

For *EL-Wikidata*, we search each mention m of an entity e in Wikidata using the Wikidata search API. This API returns a ranked list of n candidate Wikidata entities most related to the mention: $C(m) = [c_0, c_1, ..., c_{n-1}]$, c_0 being the best candidate. We give each candidate a score $\mathrm{s_{el}}$, corresponding to its index in $C(m)$

$$\mathrm{s_{el}}(m, c_i) = \begin{cases} i & \text{if } c_i \in C(m), \\ n+1 & \text{otherwise.} \end{cases} \tag{2}$$

To aggregate the candidates for all the mentions of an entity, we sum the $\mathrm{s_{el}}(m, c_i)$

$$\mathrm{s_{el}}(c_i) = \sum_{m \in e} \mathrm{s_{el}}(m, c_i). \tag{3}$$

The ranking is obtained by sorting the scores in ascending order, the first candidate (with the lower score) being the best.

EL-Wikipedia follows the same principle as *EL-Wikidata*, replacing the Wikidata search API by the Wikipedia one.

5.2 Metrics

Metrics to evaluate an end-to-end IE pipeline is 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 pipeline from the entity perspective makes sense. However, comparing one true entity with a predicted one is nontrivial because they can contain different mentions (no exact intersection) or mentions that are nearly identical but not equal (differences in boundaries, for example).

NER F1 The NER is the only module working with entity mentions. Similarly to previous works (e.g., Zhong and Chen [41]), we consider a predicted mention to be correct if its boundaries and type are the same as the ones of a ground truth mention. We use the micro aggregation for entity types to compute the F1 score[11].

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

[11]As a side note, F1 micro is equal to the accuracy in the case of a single label prediction.

Table 3: Evaluation of the PNA baseline and other approaches on the `dev` split of Linked-DocRED. For Entity F1 and Relation F1, the soft metric is displayed along with the hard aggregation in parenthesis. ATLOP [42] has access to ground truth entities and coreferences during evaluation.

	NER - Coref - RE						Entity-Linking				
Method	Mention F1 ↑	Coref. B^3 ↑	Entity F1 ↑		Relation F1 ↑		Method	Hit@1 ↑	Hit@5 ↑	NF ↓	MR ↓
Verlinden et al. [32]	-	-	-	(71.8)	-	(25.7)	-	-	-	-	-
ATLOP [42]	-	-	-	-	63.4	(63.4)	-	-	-	-	-
Ground Truth	-	-	-	-	-	-	EL-Wikipedia	52.3	61.7	32.1	2.1
							EL-Wikidata	59.0	68.5	26.3	1.7
PNA (ours)	77.2	80.4	83.9	(82.9)	48.9	(41.1)	EL-Wikipedia	46.0	53.9	40.8	2.1
							EL-Wikidata	51.1	59.1	36.2	1.7

Coref. B^3 To evaluate coreferences, we use the B^3 metric [1], which is used to evaluate clustering. This metric, among others, is recommended to evaluate coreference resolution models [27, 24, 37].

Entity F1 To provide a global metric to evaluate the extraction of entities (taking into account NER and coreferences), we recommend using the soft entity-level metric proposed by Zaporojets et al. [37].

Relation F1 Comparing a predicted relation to a ground truth relation is not trivial. Indeed, it is particularly difficult to compare entities, as they are clusters of mentions, clusters that can be both incomplete and impure. One solution can be to discard all predicted entities that are not identical to gold entities. But it does not seem fair to eliminate an entity and all its relations if it is missing only one coreference. Fortunately, Zaporojets et al. [37] proposed a soft entity-level Relation F1 score, which tackles this problem. In a nutshell, it compares the relations at a mention level, checking that both predicted mentions correspond to gold entities and that there is a relation between them. Then it aggregates the results at the entity level.

In Table 3, for Entity F1 and Relation F1, we show the soft metric but we also display the hard aggregation in parenthesis (defined by Zaporojets et al. [37]), to compare with other approaches.

Entity-Linking 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. The proportion of entities where the entity-linker does not find the correct resource.
- Mean Rank. For found entities only, the average rank where the correct resource is found.

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. *EL-Wikidata* and *EL-Wikipedia* return an ordered list of candidates for each predicted entity. For each mention in the gold entities, we find the corresponding predicted mention, if it exists, to get the ordered list of candidates associated with the mention. We then merge the candidates of all the linked mentions for each gold entity, using the same principle described for *EL-Wikidata*.

During entity-linking evaluation, NUM or TIME entities are ignored as they are not disambiguated.

Finally, Linked-DocRED, like all IE datasets, is incomplete: there is no guarantee that all entities and relations have been labeled. Precision measures have to be taken with a grain of salt, as it is not always clear if a prediction is wrong or if it corresponds to a missing entity, coreference, or relation. It impacts NER F1, Coref. B^3, Entity F1 and Relation F1. In practice, our proposed baseline is very balanced between Precision and Recall, which is a reassuring behavior.

5.3 Results

The evaluation results are shown in Table 3. We also display the results of an IE pipeline from Verlinden et al. [32], and the RE model ATLOP [42] with ground truth entities and coreferences. All methods are trained on the `train` split of Linked-DocRED and evaluated on its `dev` split. For Entity F1 and Relation F1, we show the soft metric and the hard metric (in parenthesis, to provide a comparison with [32, 42]).

Firstly, the Mention F1, Coref. B^3, and Entity F1, are superior to 75%, which is in the range of what is currently state-of-the-art for DocRED [32]. Compared to Verlinden et al. [32], 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. [28] on the DWIE dataset.

The performance of our baseline in RE is relatively low when we look at Table 3. There is obviously some error cascading, as the NER and the coreference resolver are imperfect. In fact, if we compare to ATLOP [42] with ground truth entities and coreference, the difference in soft Relation F1 is 14.5 points (23% of difference). It demonstrates that a full document-level relation extraction pipeline is a very challenging task.

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 (relations `rdfs:label` or `rdfs:aliases`), 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 1/3 of entities are wrongly disambiguated (Not Found), and only $50-60\%$ of entities are correctly disambiguated with the first match (Hit@1). It is clear that the entity-linking task is challenging, in particular when you take into account an imperfect entity and coreference extraction.

6 Conclusion and Future Work

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 metric to finalize the definition of a complete benchmark for end-to-end IE pipeline evaluation.

In the future, we plan to explore and improve information extraction pipelines and particularly compare the performance of explicit and implicit knowledge representations. We further envision to *close the loop* of information extraction, that is, benefit from the already extracted knowledge to improve the performance of the IE pipeline, which will in turn, enrich the extracted knowledge graph.

Acknownledgements

This work is supported by Alteca and the French Association for Research and Technology (ANRT) under CIFRE Ph.D. fellowship n°2021/0851. We thank the anonymous reviewers for their careful reading and insightful comments.

References

[1] 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*, pages 563–566, Granada, Spain, 1998. European Language Resources Association.

[2] 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*, pages 2670–2676, Hyderabad, India, 2007. Morgan Kaufmann Publishers Inc.

[3] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer, December 2020. URL `http://arxiv.org/abs/2004.05150`. arXiv:2004.05150 [cs].

[4] Xiaojun Chen, Shengbin Jia, and Yang Xiang. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 141:112948–112948, March 2020. doi: 10.1016/j.eswa.2019.112948. Publisher: Pergamon.

[5] 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*, pages 2819–2831, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.249. URL `https://aclanthology.org/2021.findings-acl.249`.

[6] 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*, pages 4925–4936, Hong Kong, China, 2019. Association for Computational Linguistics. ISBN 978-1-950737-90-1. doi: 10.18653/v1/d19-1498. URL `https://aclanthology.org/D19-1498`.

[7] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20 (1):37–46, 1960.

[8] 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*, volume 1, pages 4171–4186, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics. ISBN 978-1-950737-13-0. doi: 10.18653/V1/N19-1423.

[9] 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*, pages 967–977, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.84. URL `https://aclanthology.org/2020.coling-main.84`.

[10] 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*, pages 3448–3452, Miyazaki, Japan, 2018. European Language Resources Association. ISBN 979-10-95546-00-9. URL `https://aclanthology.org/L18-1544`.

[11] 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*, pages 1625–1628, Toronto, ON, Canada, 2010. Association for Computing Machinery.

[12] 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*, pages 6250–6255, Hong Kong, China, 2019. Association for Computational Linguistics. ISBN 978-1-950737-90-1. doi: 10.18653/v1/d19-1649. URL https://aclanthology.org/D19-1649.

[13] Pierre-Yves Genest, Pierre-Edouard Portier, Elöd 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*, page 11, Atlanta, USA, October 2022. ACM. doi: 10.1145/3511808.3557422. URL https://hal.science/hal-03858264.

[14] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. A Survey on Knowledge Graph-Based Recommender Systems. *IEEE Transactions on Knowledge and Data Engineering*, 34 (8):3549–3568, August 2022. doi: 10.1109/tkde.2020.3028705. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

[15] 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*, pages 4803–4809, Brussels, Belgium, 2018. Association for Computational Linguistics. ISBN 978-1-948087-84-1. doi: 10.18653/v1/d18-1514.

[16] 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*, volume 2, pages 998–1004, Online, June 2021. Association for Computational Linguistics. ISBN 978-1-954085-52-7. doi: 10.18653/v1/2021.acl-short.126. URL https://arxiv.org/abs/2106.01793v1.

[17] 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*, pages 105–113, Melbourne, Australia, January 2019. Association for Computing Machinery, Inc. ISBN 978-1-4503-5940-5. doi: 10.1145/3289600.3290956.

[18] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707, February 1966.

[19] 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. *Database*, 2016:baw068, January 2016. ISSN 1758-0463. doi: 10.1093/database/baw068. URL https://doi.org/10.1093/database/baw068.

[20] 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*, pages 3219–3232, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1360. URL https://aclanthology.org/D18-1360.

[21] 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*, volume 1, pages 3036–3046, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. ISBN 978-1-950737-13-0. doi: 10.18653/v1/n19-1308.

[22] Pablo N. 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*, pages 1–8, Graz, Austria, 2011. Association for Computing Machinery. ISBN 978-1-4503-0621-8. doi: 10.1145/2063518.2063519.

[23] 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*, pages 749–758, Hong Kong, China, 2019. Association for Computational Linguistics. ISBN 978-1-950737-90-1. doi: 10.18653/v1/d19-1069. URL https://aclanthology.org/D19-1069.

[24] 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)*, pages 632–642, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1060. URL https://aclanthology.org/P16-1060.

[25] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, March 1970. ISSN 0022-2836. doi: 10.1016/0022-2836(70)90057-4. URL https://www.sciencedirect.com/science/article/pii/0022283670900574.

[26] 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*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. ISBN 978-1-937284-96-1. doi: 10.3115/v1/d14-1162. URL `https://aclanthology.org/D14-1162`.

[27] Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2014:30–35, June 2014. ISSN 0736-587X. doi: 10.3115/v1/P14-2006. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5667668/`.

[28] Maxime Prieur, Cédric Du Mouza, Guillaume Gadek, and Bruno Grilhères. Peuplement de base de connaissances, liage dynamique et système end-to-end. *Revue des Nouvelles Technologies de l'Information*, Extraction et Gestion des Connaissances, RNTI-E-39:281–288, 2023. URL `https://hal.science/hal-03887658`.

[29] 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)*, volume 6323 LNAI, pages 148–163, Berlin, Heidelberg, 2010. Springer. ISBN 3-642-15938-9. doi: 10.1007/978-3-642-15939-8_10. Issue: PART 3.

[30] 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)*, pages 109–126, Gaithersburg, MD, USA, 1994. DIANE Publishing Company.

[31] 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*, pages 5357–5366, Punta Cana, Dominican Republic, December 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.435. URL `https://aclanthology.org/2021.emnlp-main.435`.

[32] 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, July 2021. URL `http://arxiv.org/abs/2107.02286`. arXiv:2107.02286 [cs].

[33] 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*, pages 5784–5789, Hong Kong, China, 2019. Association for Computational Linguistics. ISBN 978-1-950737-90-1. doi: 10.18653/v1/d19-1585.

[34] Xinyi Wang, Zitao Wang, Weijian Sun, and Wei Hu. Enhancing Document-Level Relation Extraction by Entity Knowledge Injection. In Ulrike Sattler, Aidan Hogan, Maria Keet, Valentina Presutti, João Paulo A. Almeida, Hideaki Takeda, Pierre Monnin, Giuseppe Pirrò, and Claudia d'Amato, editors, *The Semantic Web – ISWC 2022*, Lecture Notes in Computer Science, pages 39–56, Cham, 2022. Springer International Publishing. ISBN 978-3-031-19433-7. doi: 10.1007/978-3-031-19433-7_3.

[35] Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22 (158):209–212, 1927. doi: 10.1080/01621459.1927.10502953. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.1927.10502953`.

[36] 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*, pages 764–777, Florence, Italy, 2019. Association for Computational Linguistics. ISBN 978-1-950737-48-2. doi: 10.18653/v1/p19-1074.

[37] Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. DWIE: An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*, 58(4):102563, July 2021. ISSN 03064573. doi: 10.1016/j.ipm.2021.102563. URL `https://linkinghub.elsevier.com/retrieve/pii/S0306457321000662`.

[38] 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*, pages 1630–1640, Online, September 2020. Association for Computational Linguistics (ACL). ISBN 978-1-952148-60-6. doi: 10.18653/v1/2020.emnlp-main.127. URL `https://arxiv.org/abs/2009.13752v1`.

[39] 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*, volume 1, pages 3016–3025, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. ISBN 978-1-950737-13-0. doi: 10.18653/v1/n19-1306. URL `https://aclanthology.org/N19-1306`.

[40] 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*, pages 1630–1641, Barcelona, Spain, January 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.coling-main.143. URL `https://aclanthology.org/2020.coling-main.143`.

[41] 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*, pages 50–61, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.5.

[42] 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*, volume 35, pages 14612–14620, Online, May 2021. AAAI Press. doi: 10.1609/aaai.v35i16.17717. URL `https://ojs.aaai.org/index.php/AAAI/article/view/17717`. Number: 16.