論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Information Extraction Beyond Sentence Boundary
著者(和文)	MAYoumi
Author(English)	Youmi Ma
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12916号, 授与年月日:2024年9月20日, 学位の種別:課程博士, 審査員:岡崎 直観,徳永 健伸,宮崎 純,村田 剛志,金崎 朝子
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12916号, Conferred date:2024/9/20, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Doctoral Dissertation

Information Extraction Beyond Sentence Boundary

Youmi Ma

August 5, 2024

Artificial Intelligence Course
Department of Computer Science
School of Computing
Tokyo Institute of Technology

A Doctoral Dissertation
submitted to the School of Computing,
Tokyo Institute of Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Youmi Ma

Thesis Committee:

Professor Naoaki Okazaki	(Supervisor)
Professor Takenobu Tokunaga	(Co-supervisor)
Professor Jun Miyazaki	(Co-supervisor)
Professor Tsuyoshi Murata	(Co-supervisor)
Associate Professor Asako Kanezaki	(Co-supervisor)

Information Extraction Beyond Sentence Boundary *

Youmi Ma

Abstract

Information Extraction (IE) is the task of extracting structured information from unstructured texts. The collected structured information can be transformed into a Knowledge Base (KB), serving as a valuable assistant for human decision-making. Relation Extraction (RE), an important subfield of IE aiming at extracting relation triples in the form of (subject, relation, object), is closely related to the automatic construction of graph-shaped KB. While classical RE is a sentence-level task, recent studies have pointed out that sentence-level RE is impractical, as relations can hold document-wise, i.e., beyond sentence boundaries. The task, Document-level Relation Extraction (DocRE), is thus proposed to encourage extracting both intra- and inter-sentence relation triples.

However, due to the complexity of DocRE, human-annotated supervisory signals of high quality are limited and difficult to expand. This results in two challenges of DocRE regarding human annotations. Firstly, the limited human annotations are not fully utilized to train a better DocRE model. Specifically, evidence annotations – a set of sentences necessary to identify the relation between an entity pair – are provided alongside relation annotations but are not used to train a DocRE model. Secondly, there is no methodology that enables efficient construction of a DocRE dataset in a new language. While there is a demand for automatically populating KB for each language, datasets supporting the training of DocRE models are limited to only two or three languages. This study explores ways to address the aforementioned challenges. The core idea is to better utilize existing language resources with human annotations to help model construction and dataset construction.

*Doctoral Dissertation, School of Computing
Tokyo Institute of Technology, August 5, 2024.

For model construction, the goal is to obtain a DocRE model with high performance. To this end, this study proposes a training strategy named **D**ocument-level **R**elation **E**xtraction with **E**vidence-guided **A**ttention **M**echanism (DREEAM). DREEAM incorporates evidence supervisory signals into the parameter updates of DocRE models. When deciding the relation(s) between a (subject, object) entity pair, models are trained to pay more attention to sentences marked as evidence by human annotators. The study further proposes an approach to assign pseudo-evidence to data without evidence annotations. These data, once assigned with pseudo-evidence, are also used to train an improved DocRE model. The two proposals altogether yield a state-of-the-art DocRE model on multiple benchmarks. Notably, DREEAM is memory-efficient, reducing memory usage during inference to 30% of that required by existing methods. DREEAM also enhances explainability compared to the baseline method by guiding attention with evidence.

For dataset construction, the goal is to obtain a dataset in a language without DocRE language resources, with reduced annotation costs. To this end, this study selects Japanese as the target language and conducts cross-lingual projection from existing language resources in English. A machine translator translates the documents in the English dataset while simultaneously projecting the entity label spans. However, models trained on the translated dataset failed to extract many relation triples from raw Japanese texts. Having witnessed the failure of the translated dataset, this study further proposes a semi-automatic method to employ the dataset as an assistant to human annotations. The machine-human collaborative scheme requires annotators to revise recommendations provided by DocRE models trained on the translated dataset. Compared with existing annotation approaches, the proposed scheme reduces the number of human annotation steps to more than half. As a result, JacRED, the first general-purpose **Ja**panese Do**c**ument-level **R**elation **E**xtraction **D**ataset, is published along with the new annotation scheme. Notably, while 45% of the relation triples in existing English language resources can be extracted from a single sentence, the percentage of intra-sentence relation triples is reduced to 33% in JacRED. The fact suggests that JacRED is more aligned with the objective set by DocRE, focusing on cross-sentence relation extractions. Experiment results have confirmed that JacRED's quality is superior to the translated dataset. When benchmarking with JacRED, DREEAM, the method proposed in this study, still ranks first among all exist-

ing methods, demonstrating its superiority in extracting relations and retrieving evidence. Large Language Models such as GPT-3.5 or GPT-4 are not as good as supervised methods in DocRE. Additionally, JacRED, together with the English DocRE dataset, enables the evaluation of cross-lingual DocRE.

This study contributes to the Natural Language Processing (NLP) field by offering methods and insights that enhance the accuracy and expand the applications of DocRE, eventually enhancing knowledge base completion. As knowledge bases are attracting increasing attention in improving the reliability of generative AIs based on large language models, enhancing knowledge base completion benefits the research field in developing responsible and reliable AIs.

From the perspective of societal applications, enhancing knowledge base completion will make organizing and managing unstructured data easier. The need to structuralize data and organize the information not only resides in data publicly available on the World Wide Web but also in confidential data in administrations or companies. This study, therefore, provides advancements for both the NLP field and broader societal applications.

Keywords:

Information Extraction, Document-level Relation Extraction, Evidence Retrieval, Cross-lingual Projection, Dataset Construction

Acknowledgements

本研究を遂行するのにあたり，多くの方々からご指導およびご支援を賜りました．この場を借りて心より感謝を申し上げます．

主指導教員の岡崎直観先生には，本研究の立案から論文執筆まで，所々から丁寧かつ的確なご指導を賜りました．修士課程からの5年間，ご多忙であるにも関わらず，いつも親身に相談に乗ってくださった．修士段階では丁寧な指南を，博士段階では的確な指摘をいただいたおかげて，研究者としての成長を実感できました．直接なご指導以外にも，教育活動や研究活動に限らず，学会活動にも真摯に向き合う先生の姿勢から薫陶を受けました．入学時には立ち上げたばかりの若い研究室でしたが，今では国内外から注目を集めるほど名高くなりしました．研究室の急成長期に入学し，その成長を見届けられたことを誇らしく思います．また，留学生活の序盤では，研究生活だけでなく私生活でも不運が続き，ご心配をおかけしました．博士課程の後半では，自分の研究テーマと時代の流れとの食い違いに圧倒され，落ち込む時期もありました．これらの困難を乗り越え，博士課程まで完走できたのは，間違いなく，温かい応援と手厚いサポートをくださった岡崎先生のおかげです．先生には感謝してもしきれません．

徳永健伸教授には，同じ自然言語処理の先生として，博士論文の審査のみならず，学会等でもお世話になりました．中間発表では，根拠文獲得は関係抽出の副産物だけではなく，関係抽出器の説明可能性を向上する手段にもなるとご助言いただきました．予備審査では，二番目の研究の難解な部分や，議論が不足していた部分についてご指摘いただきました．徳永先生のご助言は，本研究の完成度を向上させる上で，大いに助けとなりました．

宮崎純教授には，提案手法の適用範囲や，実社会応用を見据えた際に解決しなければならない課題についてコメントいただきました．宮崎先生のコメントは，本研究の実社会応用における立ち位置を客観的に評価する助けとなりました．

村田剛志教授には，根拠文ラベルの妥当性や，提案したアノテーション手法のコストについてご指摘いただきました．特に一番目の研究は根拠文に依存が強いため，根拠文の予測結果だけでなく，ラベル自体の分析も重要であると気づき，それを原稿に追記しました．村田先生のご指摘は，本研究のサウンドネスを向上するのに必要不可欠でした．

金崎朝子准教授には，一番目の研究における学生モデルと教師モデルの関係性

や，二番目の研究における提案手法の立ち位置に関するコメントをいただきました．金崎先生のコメントは，本研究の学術・技術的な貢献を深く考え直すきっかけとなりました．

研究支援員の中川恵理子さん，小西由希子さん，雲財祐子さん，古谷奈緒子さんからは，日頃の研究生活のご支援を賜りました． 中川さんには，留学生活の序盤で心細かった頃に，研究環境の整備をサポートいただきました． 小西さんには，修了手続きやリサーチ・アシスタントの勤怠管理を始め，多くの事務手続きでお世話になりました．また，研究室行事の記録としての写真撮影も何回か代行してくださいました． 雲財さんには，計算資源の課金管理や論文出版，備品購入などでお世話になりました．また，日頃から親切にしてくださり，いつも雑談を楽しませていただきました． 古谷さんには，出張手続きで大変お世話になりました．特に海外出張するためのビザ申請に関して，何度も相談に乗っていただき，多くの助言と助力を賜りました．

NTT人間情報研究所の西田京介さんと西田光甫さんには，研究インターンでお世話になりました． 西田光甫さんには，インターンのメンターとして丁寧なご指導を賜りました． 西田京介さんには，インターンの統括責任者として，研究が行き詰まるときにアドバイスをいただきました． 普段とは異なる研究テーマを取り組むことにより，情報抽出に拘っていた私の視野を広げることができ，本研究の発案に至りました． インターンのテーマを提案してくださったお二方に感謝いたします．

I want to express my deepest thanks to Dr. Bhushan Kotnis, Dr. Carolin Lawrence, and Prof. Goran Glavaš. It is my honor to have an 8-month collaboration project with them, from which I learned a lot. Although the project is not directly related to this dissertation, the experience has enhanced my research skills, helping me to successfully complete my doctoral research. Dr. Bhushan Kotnis mentored the whole collaboration project with great patience. His excellent coding skills greatly impressed me and taught me a lot about code management. Dr. Carolin Lawrence provided all kinds of support to ensure the project could be carried out smoothly. Before submitting the paper, she carefully reviewed all the details and provided valuable suggestions. Pro. Goran Glavaš joined the weakly meeting with great passion as an advisor. He is visionary and foresees the challenges and chances of the research project. Even outside the project, he has shown great kindness in sharing his career path with me and never hesitated to lend me a hand whenever necessary. I am looking forward to meeting all of them and expressing my thanks in person in the future.

岡崎研究室の研究員や，博士課程を卒業された方々の背中を追ってここまで来れました． 高瀬翔元助教には，キャリア設計において多大なるお力添えをいただきました． ご在籍中はインターンに関する貴重なアドバイスをいただき，インターン先選定の決め手となりました． ご転職された後も，ご勤務先にお邪魔させていただい

た時に，わざわざお時間を割いて話を聞かせてくださいました．進路を決めた後にも，温かくて心強いお言葉を賜りました． また，研究活動においても，高瀬さんの効率的な研究の回し方が指針となっており，それに準じて研究サイクルを定期的に見直しています． 2021年度修了生で，研究員として在籍されていた平岡達也さんには，修士段階の研究をメンターリングいただきました． 研究の方向性調整や研究進捗の管理，論文の添削まで，所々でご指導を賜りました． 卒業された後にも，研究やキャリアに関する相談をさせていただいており，いつも応援してくださいました． YANS委員の後継を私に指名してくれましたことも大変嬉しくて，光栄に思います． 初志を貫徹し，徐々に分野の第一人者となっていく平岡さんの姿を見て，同じく基礎研究を取り組む者として，鼓舞を受けました． まだ一人前になれたかどうかは微妙なところではありますが，ご厚意に恥じぬように精進したいと思います．特別研究員の金子正弘さんとは，共に卒論生のメンターとして研究させていただきました． 金子さんのエネルギッシュな研究スタイルには度々感服しており，共に研究ができることを誇らしく思います． また，ペット飼育などに関する雑談も楽しませていただきました， 2022年度修了生で，今も研究員として在籍されている丹羽彩奈さんには，いつも親切にしていただきました． コロナ禍の中でも，コロナが落ち着いた後でも，積極的に対外活動を行い，多くの人と繋がり，コミュニティに貢献しようとしている丹羽さんを尊敬しています． また，就職活動の相談にも，何度も飽きずに乗ってくださいました． 丹羽さんや，丹羽さんからの女子会やDの会へのお誘いがなければ，私は孤独な5年間を過ごしていたのに間違いありません． 女性研究者の先輩として，丹羽さんが同じ研究室にいることを心から幸いに思います．卒業後も仲良くしていただけると嬉しいです． 2023年度修了生で，今も研究員として在籍されている水木栄さんには，研究テーマの相談や大規模言語モデルプロジェクト「Swallow」でお世話になりました． 水木さんの抜群の思考力や，緻密な計画を立てるプランニング能力には脱帽です． それに加え，常に勤勉であるところも敬服しています． 仮説を立て，実験で検証する研究プロセスは，水木さんの見様見真似です． また，日本語のお手本とさせていただいた時期もありました． I want to thank Dr. Ao Liu, who graduated from our lab in 2023. As we are both international students from China, I truly appreciate your company. 2023年度修了生の飯田大貴さんには，定例のResearch Seminarの発表で度々鋭いご指摘をいただきました． また，雑談にも付き合ってくださり，楽しい時間を過ごすことができました．

岡崎研究室の皆さんには，定例セミナーでの研究議論から日頃の交流まで，様々な面でお世話になりました． I want to appreciate my deepest thanks to An Wang, who collaborated with me on both studies included in this dissertation. He is always kind, optimistic, and open to discussions. The dissertation could not have been completed without his advice. 服部翔さん，綿祐貴さんとは，メンターとして共同研究をさせていただきました． メンターとして大変未熟で，迷惑ばかりかけていた

かもしれませんが，研究を楽しく続けられると嬉しいです． 郭瑄瑜さんには，共同研究の他に，私生活でも仲良くしていただきました． 舜夏海さんには，入学したばかりの頃，研究や生活の面倒を見てくださいました． 歳の近い先輩として，精神的に大きく支えられました． また，ご卒業された後でも，学会などで雑談に付き合っていただきました． My thanks go to Zhishen Yang and Sangwhan Moon for their kindness and support. I have enjoyed both the welcome party they held for me and our daily chats. 村岡雅康さんには，研究発表会や論文輪読会などの場で，多くの議論をさせていただきました． 積極的に質問する村岡さんの姿勢を，いつも見習いたいと思っています． また，インターンの相談にも乗ってくださいました． 吉川和さんには，同じ博士課程の女子学生として，仲良くしていただきました． また，就職活動の相談に乗ってくださっただけでなく，面談の機会まで設けていただき，大変お世話になりました． 吉川さんのご親切さにはいつも助けられてばかりです． My great thanks go to Marco Cognetta, who has always been kind and energetic. The paper would not have been accepted to EACL without your proofreading. Also, thanks so much for naming the method “DREEAM”; I really like it. I also want to thank Erick Mendieta Molina and Vijay Daultani for providing valuable advice about career planning. 古山翔太さん，小池隆斗さん，前田航希さん，大井聖也さんをはじめ，他の研究室のメンバーには，研究議論や日常の雑談でお世話になりました． 日本語や英語の練習も兼ねて，付き合っていただきありがとうございました． 全ての方々のお名前を列挙しきれないことをお詫びいたします． 岡崎研究室の皆さんは優秀な方ばかりで，共に過ごす時間は楽しく，多くの刺激を受けることができました． 皆さんと同じ研究室に在籍できたことを誇らしく思います． 留学生の私を優しく受け入れていただき，本当にありがとうございました．

高校や大学で出会った友人たちは，相変わらず雑談や愚痴の相手になってくれました． 卒業後は世界各地で暮らすことになり，コロナ禍でさらに会うのが難しくなりましたが，それでも連絡を取り合ってくれていることに感謝しています．

最後に，修士・博士一貫コースで日本の大学に留学することに躊躇いもなく応援してくれた父，母，姉に感謝します． 思いもよらぬ激変も少なからずありましたが，念願の博士号はなんとか滞りなく取得できそうです． これからも研究の道を選んだ私を，これまで通りにサポートしてくれると嬉しいです．そして作業中はいつも隣で寝転がってくれるソマリ猫のシルバンに感謝します．

Contents

List of Figures

List of Tables

1 Introduction

1.1 Information Extraction

The World Wide Web enables storing and sharing texts over the Internet, including but not limited to newswire articles, reports, and social networking services. These texts contain information that may be of interest to others. For example, financial traders are interested in information from newswire articles to decide which financial asset to buy or sell [21]; doctors are interested in information from medical reports to determine the appropriate drugs for their patients [57].

However, it is time-consuming to manually extract the information of interest from a large number of texts, as they are highly unstructured and in free form, written by various individuals for different purposes. This induces the need for automatic **Information Extraction** (**IE**), where a (pre-specified) sort of information is extracted from unstructured natural language texts [34].

IE has been an essential subfield of **Natural Language Processing** (**NLP**) research, with a history dating back to the 1980s [35]. In the early days, IE usually involved filling a pre-defined template, typically solved by rule-based approaches such as pattern matching [3, 36]. Entering the 21st century, Doddington et al. [29] reduced the complex template-filling task to extracting text spans from the original document. They decompose IE into 3 subtasks, as shown in Figure 1.1.

1. **N**amed **E**ntity **R**ecognition (**NER**, [22, 45, 84, 90]) for detecting every text span that represents an entity of a pre-defined type, e.g., *Steve Jobs* ← `PERSON`.

2. **R**elation **E**xtraction (**RE**, [67, 72, 83, 115]) for detecting every pair of text spans (s, o)[1] that a pre-defined relation `r` holds between s and o, e.g., (*Steve Jobs*, `work_for`, *Apple*).

[1] s stands for *subject* and o stands for *object*.

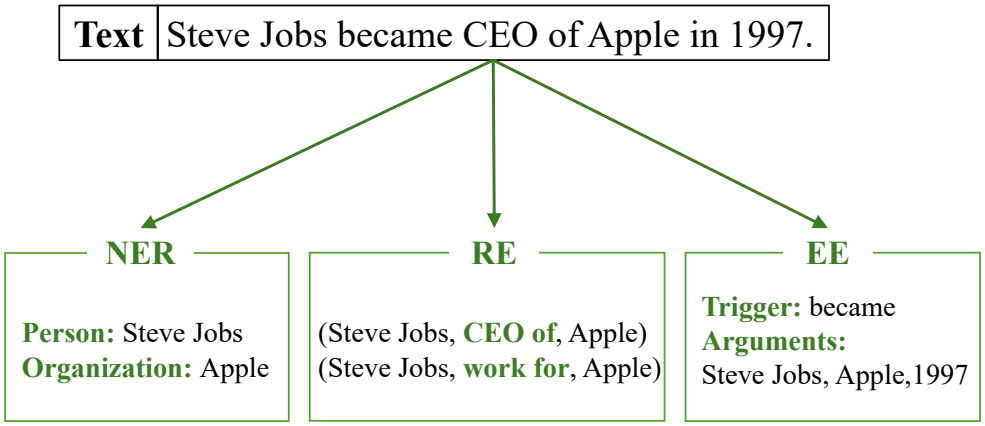


Figure 1.1: Examples of typical Information Extraction tasks.

3. **E**vent **E**xtraction (**EE**, [63, 97, 109]) for detecting a series of text spans that corresponds to the `trigger` and its associated `arguments` of an event, e.g., [`Trigger`: *became*, `Arguments`: (*Steve Jobs*, *Apple*, *1997*)].

The same work also published a corpus with human-annotated labels for each subtask, enabling the supervised training of machine-learning models. Recent IE research follows this scope of task definition while trying to bring the settings closer to reality [33, 112].

Among all these subtasks, RE contributes to another task named **K**nowledge **B**ase **P**opulation (**KBP**), aiming at gathering structured information extracted from different sources [47]. As in Figure 1.2, information scattering among multiple documents is extracted with IE models and combined into a structured **K**nowledge **B**ase (**KB**), typically shaped as a directed graph. As queryable, fast, and reliable data sources, KBs have been widely used to benefit downstream NLP tasks such as question answering [25, 38].

Recently, KBs are attracting even more attention, owing to the emergence of **L**arge **L**anguage **M**odels (**LLM**) [76]. These models are powerful natural language generators but suffer from low explainability and reliability due to hallucinations. LLMs may generate inaccurate, misleading, or inconsistent responses not supported by factual information [42]. Researchers have been discussing KBs' potential to mitigate hallucinations. By incorporating knowledge retrieved from KBs into the input of LLMs, one can observe an improved performance on question-answering tasks [1, 6, 61]. KBs can also act as a verifier to assess the

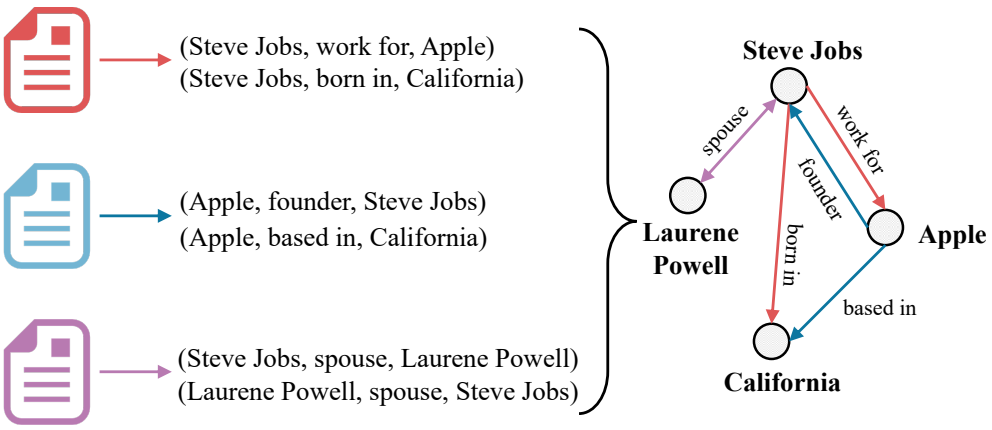


Figure 1.2: Example of knowledge base gathering extractions from multiple documents.

correctness of outputs generated by LLMs [51, 65]. Moreover, reasoning on the graph structure of KBs improves the explainability of LLMs, as the reasoning path can be explicitly grounded to the graph [68, 99]. It is notable that RE, as a subtask of IE, is a crucial step for automatically constructing and enhancing knowledge bases, from which we witness the importance of IE throughout the history of NLP.

This study focuses on RE, the task that can help populate KBs and benefit downstream NLP applications such as LLMs. Typically, RE is a sentence-level task that aims at extracting triple(s) with pre-defined relation type(s) from a sentence (Figure 1.1,[29, 82, 120]). However, existing studies have pointed out that sentence-level RE is over-simplified, as **relations can hold beyond sentence boundaries**, i.e., between entities in different sentences [18, 59, 112]. As shown in Figure 1.3, an RE model practical for KBP should be able to extract both intra- and inter- sentence triples. Such a need to develop RE models ready for practical use induces a variant of RE tasks, namely **Doc**ument-level **R**elation **E**xtraction (**DocRE**), which is the target of this study.

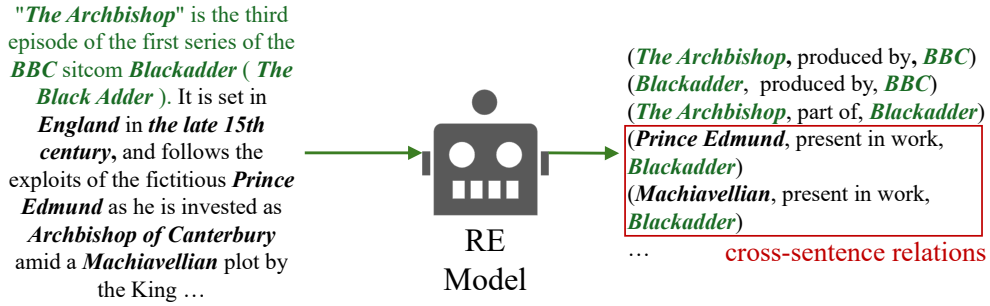


Figure 1.3: Illustration of a DocRE model extracting intra-sentence and cross-sentence relation triples. Words in the first and the second sentences are marked in green and black, respectively.

1.2 Document-Level Relation Extraction and the Challenges

DocRE aims to identify *all* semantic relationships between entities in a document [112]. The task promotes RE to a more practical setting, where relations can reside between entity pairs *document-wise*, i.e., within and beyond the sentence boundary, as in Figure 1.3. For this reason, DocRE has been recognized as a more challenging task compared with its sentence-level counterpart [77, 94, 112]. DocRE is also worth spotlighting because the task showcases how models comprehend long text [14, 91, 114]. Notably, although LLMs have "solved" various NLP tasks such as text summarization with performance defeating supervisedly-trained models, their performance on DocRE stays low (Section 5.3, [60, 98]). Therefore, DocRE deserves attention even in the era of LLMs.

The development of DocRE is accompanied by a shortage of supervised data. Compared with sentence-level RE, DocRE is more difficult to annotate, as enumerating relationships within a document is much harder than those within a sentence. The difficulty arises not only from the increased number of entities but also from the complexity of the text, which requires more time to comprehend. As a result, collecting human annotations for DocRE is costly and time-consuming, making every human-annotated corpus precious. For the time being, a majority of DocRE datasets collect relation labels by automatic assignment using machines, without the help of human annotators [27].

Closely related to the scarcity of human-annotated data, researchers are faced

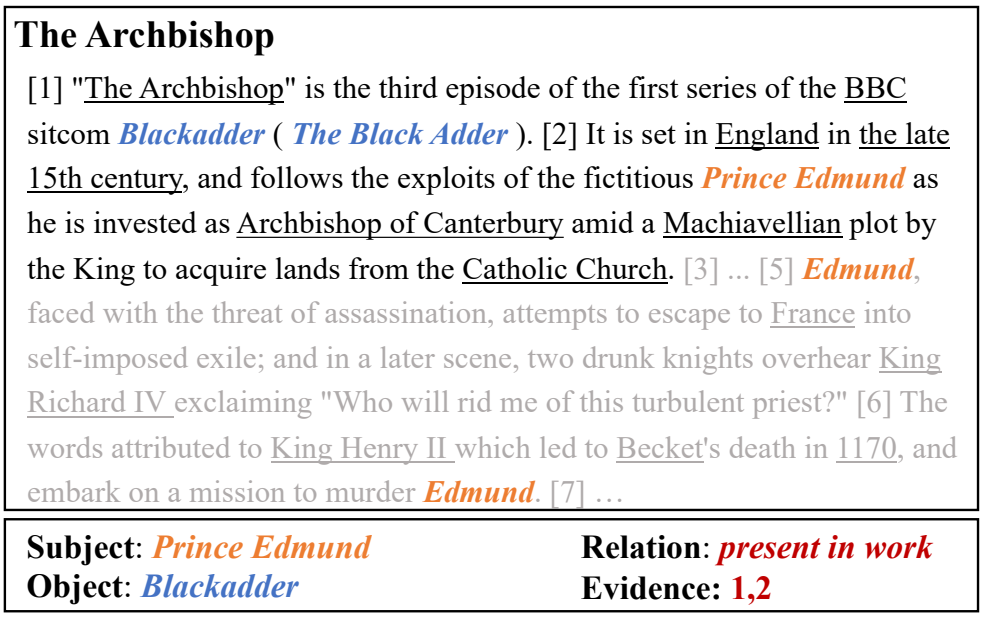


Figure 1.4: Example document and one of the relation triples from DocRED, where the i-th sentence is marked with [i] in the beginning. Mentions in bold italics are those of subjects and objects, whereas entity mentions other than subject and object are underlined.

with two **challenges** in order to solve DocRE.

1.2.1 Insufficient Usage of Human Annotations

Despite the preciousness of human annotations, the manually assigned labels have not been fully utilized for training DocRE systems. To be specific, DocRED [112], the first and most popular dataset that brings the task DocRE to public notice, collects not only the relation labels but also the supporting evidence label (*evidence* hereafter) for each relation decision.

Evidence is defined as a set of sentences necessary for humans to identify the relation between an entity pair [112]. As shown in Figure 1.4, to decide the relation label *present in work* between *Prince Edmund* and *Blackadder*, reading sentences 1 and 2 should be sufficient. Although sentences 5 and 6 also mention the subject, they are irrelevant to the relation decision. Evidence of the relation triple (*Prince Edmund*, *present in work*, *Blackadder*) is thus sentences 1 and 2.

Dataset	Lang.	# Triples	# Docs.	Avg. # Toks.	# Rels.	Evi.
DocRED [112]	*en.*	50,503	4,051	198.4	96	Y
Re-DocRED [89]	*en.*	120,664	4,053	198.4	96	N
HacRED [18]	*zh.*	56,798	7,731	122.6	26	N
HistRED [111]	*kr.*	9,965	5,816	100.6	20	Y

Table 1.1: Statistics of existing DocRE datasets. Column **Evi.** shows whether each dataset annotates evidence sentences or not. Statistics for DocRED are from the human-annotated subset.

Previous studies introduce a new task called Evidence Extraction or **E**vidence **R**etrieval (**ER**), attached to DocRE [41, 43, 104, 105]. The aim of ER is to measure the ability of models to identify the evidence for each relation extraction decision. To this end, modules are specially designed to perform ER in addition to DocRE. However, the ER modules are separated from DocRE modules, with little parameter sharing or interaction [104, 105]. As a result, the supervisory signal of ER contributes little to the parameter updates of the DocRE module. In other words, no efforts are made to utilize human annotations of evidence for better extracting relation triples from documents.

1.2.2 Limited (Multilingual) Human Annotations

The scale and amount of DocRE language resources are smaller than its sentence-level counterpart. Table 1.1 summarises the existing DocRE datasets. Several datasets are excluded as they target specific domains, e.g., medical, drug, and biology. As shown in the table, English is the dominant language where two datasets have been published with the largest amount of triples[2]. Chinese and Korean datasets are also constructed, while the average number of triples per document is much smaller. These non-English datasets are constructed individually from English datasets with different label sets.

The above observations portray a gap between the availability of language resources and the practical needs: While there is a demand for automatically populating KB for each language, datasets supporting the training of DocRE

[2]To be precise, Re-DocRED is an improved version of DocRED where more relation labels are added [89, 112].

models are limited to only 2 or 3 languages. Given the hardness of collecting human annotations for DocRE, efforts should be made to either (1) train a DocRE model capable of extracting relation triples from multilingual documents or (2) develop a method for easier dataset construction in an arbitrary language.

1.3 Proposed Solutions

This study investigates ways to move one step forward in solving the abovementioned challenges. The core idea is to better utilize existing language resources with "gold", i.e., human-labeled annotations, to help (1) train a model with better performance and (2) construct a dataset with fewer costs.

1.3.1 Model Construction

To improve the usage of human annotations in training DocRE models, this study proposes a method to incorporate the supervisory signal of evidence retrieval into the parameter updates of DocRE models. The method is named **DREEAM**, short for **D**ocument-level **R**elation **E**xtraction with **E**vidence-guided **A**ttention **M**echanism[3].

Specifically, to decide the relation label between an entity pair (s, o), models are taught to pay more attention to sentences marked as evidence by human annotators. This approach aligns the behavior of DocRE models with that of humans: When asked to determine the relation between a specific entity pair based on a given document, human beings first search through the text to find relevant clues. An answer can then be made by referring to the relevant clues, which resembles the process of model predicting relation label(s).

The study further proposes an approach to assign "silver" evidence annotations to data with "silver" relation annotations. Data with silver relations and evidence annotations are further incorporated into the training process to produce a better DocRE model.

The two proposals altogether yield a state-of-the-art DocRE model on multiple benchmarks. The fact that DREEAM outperforms all its competitors demonstrates the study's success in utilizing evidence annotations to train a better model that efficiently extracts relation triples from documents.

[3]An introduction of attention mechanism is included in Section 2.1.1.

1.3.2 Dataset Construction

To promote DocRE research in languages other than English, this study proposes a method to utilize existing resources of English DocRE to construct datasets and models for non-English DocRE. Japanese is chosen as the target language, and this study eventually constructed the first **Ja**panese Do**c**ument-level **R**elation **E**xtraction **D**ataset, shortened as **JacRED**.

The study first assesses the usability of a dataset directly translated from English resources. A machine translator translates text in Re-DocRED [89] into the target language while simultaneously projecting the entity label spans. The approach has been reported effective in automatically collecting multilingual sentence-level RE datasets [40]. However, models trained on the translated dataset, although performing fairly well on the test split of the translated dataset, failed to extract many relation triples from raw texts written by native speakers in the target language.

Having witnessed that the translated dataset is not ready for use on its own, the study further proposes a semi-automatic scheme to utilize the dataset to assist human annotation. Given a raw document in the target language, DocRE models trained on the translated dataset suggest possible relation triples, based on which human annotators make modifications. The dataset constructed following this semi-automatic scheme is JacRED. This human-computer collaborative scheme helps reduce human annotation costs to more than half compared to existing work. Experimental results demonstrate that models trained with JacRED consistently outperform those trained with the translated dataset, indicating an improved dataset quality. Notably, compared with Re-DocRED, JacRED contains a higher percentage of cross-sentence relation instances, aligning better with the objective of DocRE.

Although this study specifically focuses on creating a Japanese dataset with the assistance of an English dataset, the findings arguably apply to any other language pair.

1.4 Contributions

This study contributes to the community from the following perspectives:

1. It proposes the first approach that utilizes the supervision signal of evidence

directly in the training of DocRE models.

2. It introduces the first approach to assign evidence labels to unlabeled documents automatically.

3. Combining the above approaches yields DREEAM, the state-of-the-art DocRE model, on multiple benchmarks. The proposed model is more memory-efficient, reducing memory usage during inference to 30% of that required by existing methods.

4. The proposed model is more explainable, focusing on words related to the corresponding entity pair to decide the relation label(s).

5. It publishes a ready-to-use software for extracting relation triples and their corresponding evidence from English documents.

6. It examines the validity of automatically constructed DocRE datasets via machine translation and showcases the limitations of such datasets.

7. It proposes a human-computer collaboration scheme for annotating DocRE datasets, assisted by existing datasets in another language. The proposed scheme reduces the human annotation steps to 50% of that required by existing methods.

8. It publishes JacRED, the first document-level relation extraction dataset in Japanese, establishing the foundation of Japanese DocRE.

9. Benchmarking with JacRED depicts the limitations of LLMs in solving DocRE, even with a delicately designed prompt.

10. The published dataset enables cross-lingual evaluation of DocRE models, as it shares the same domain with existing English datasets, with a transferrable relation label set.

1.5 Outline

This section provides an outline of the remaining contents.

Chapter 2: Background Knowledge

Chapter 2 walks through the background knowledge and preliminaries necessary to understand this research. This includes: (1) the attention mechanism and Transformers [93], the technologies that form the basis of modern NLP; (2) LLMs and in-context learning, the edge-cutting technologies frequently mentioned in current NLP studies, which are also utilized in this work.

Chapter 3: Preliminaries and Related Work

Chapter 3 introduces the related studies. The chapter is divided into three parts. Section 3.1 specifies the task definition and details the basic dataset utilized throughout this study. Section 3.2 introduces preliminaries and existing studies about building DocRE models, positioning this study in the family of designing a better DocRE model. Section 3.3 introduces existing studies about constructing DocRE datasets, positioning this study in the family of constructing language resources for (Doc)RE.

Chapter 4: Model Construction

Chapter 4 details the first proposal of this study about model construction, i.e., DREEAM. The aim is to incorporate supervisory signals regarding evidence to help obtain better DocRE models. The chapter first introduces the technical details and then the experiment settings and results. Analyses, including ablation studies, are also included to describe the properties of DREEAM.

Chapter 5: Dataset Construction

Chapter 5 details the second proposal of this study about dataset construction, using Japanese as the target language. The aim is to reduce the burden of human annotation to create multi-lingual and cross-lingual DocRE datasets. The chapter first details the full-automatic approach, i.e., translating English DocRE datasets into Japanese, including its limitations. Following these contents is the semi-automatic approach and an analysis of the constructed dataset, i.e., JacRED. Several experimental results about JacRED portray the reasonableness of the proposed annotation scheme and the property of the constructed dataset.

Chapter 6: Conclusion

Chapter 6 concludes the study about information extraction, especially relation extraction, beyond sentence boundaries. It first reviews the proposals and results in Chapters 3 and 4 and discusses how this study has mitigated the challenges of DocRE. Apart from the achievements, this chapter also discusses the limitations of the study and potential research directions to stimulate future research.

2 Background Knowledge

This chapter introduces the background knowledge about modern NLP, which the dissertation is based on. Specifically, the chapter walks through technologies from the attention mechanism [93] to the cutting-edge in-context learning methods [30].

2.1 Transformer

Transformer is a deep-learning model that sets the foundation for modern NLP, capable of modeling long-range dependencies between words in a sentence. The key component enabling this capability is the attention mechanism, which is also closely related to this study's proposal in Chapter 4.

2.1.1 Attention Mechanism

Proposed by Bahdanau et al. [7], the technique allows a model to focus on different parts of the input sentence dynamically. The attention mechanism can be used under different scenarios, typically cross-attention and self-attention. Cross-attention computes the alignment between the input and the output, generating the next word based on the previously seen words [7]. Self-attention computes the alignment within input words, obtaining a contextualized vector representation of each word (token) or sentence that takes all words into consideration [28]. This section introduces the latter setting, i.e., a self-attention encoder, that is closely related to this study. It is notable that while used in different scenarios, cross-attention and self-attention share the same core concept.

The following subsection explains self-attention using an example sentence "see that boy dance". Given the input sentence, the contextualized vector representation of each word is computed from a weighted sum of all word vector representations. Here, the process of computing the contextualized vector representation of "that" is showcased in Figure 2.1. Mathematically, let the word vector matrix

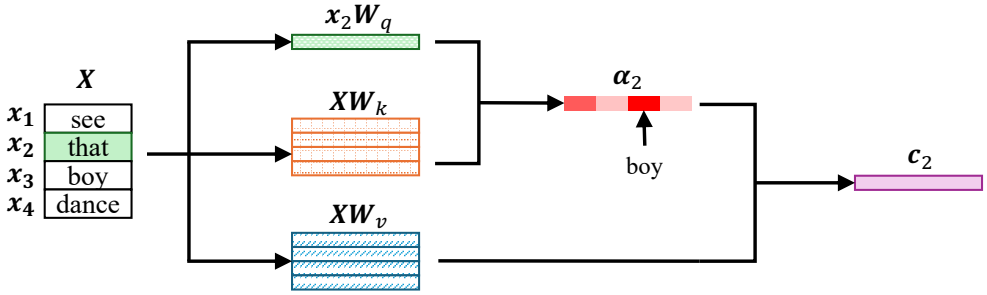


Figure 2.1: Example of attention mechanism originating from the word "that".

be $\boldsymbol{X} := [\boldsymbol{x}_1; \boldsymbol{x}_2; \boldsymbol{x}_3; \boldsymbol{x}_4] \in \mathbb{R}^{4\times d}$, where d is the dimension size of word vectors. A query, key, and value matrix is then computed as $\boldsymbol{X}\boldsymbol{W}_q, \boldsymbol{X}\boldsymbol{W}_k, \boldsymbol{X}\boldsymbol{W}_v \in \mathbb{R}^{4\times d_k}$, where $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbb{R}^{d\times d_k}$ are matrices that project word vectors into the space of being query, key, and value. The contextualized word vector of "that" is computed by querying with $\boldsymbol{x}_2\boldsymbol{W}_q$. Firstly, the importance of each word $\boldsymbol{\alpha}_2 \in \mathbb{R}^4$ are computed as:

$$\boldsymbol{\alpha}_2 = \frac{(\boldsymbol{x}_2\boldsymbol{W}_q)(\boldsymbol{X}\boldsymbol{W}_k)^\top}{\sqrt{d_k}}. \tag{2.1}$$

The contextualized word vector $\boldsymbol{c}_2$ is then computed as a weighted sum over $\boldsymbol{X}\boldsymbol{W}_v$, with weights given by $\boldsymbol{\alpha}_2$:

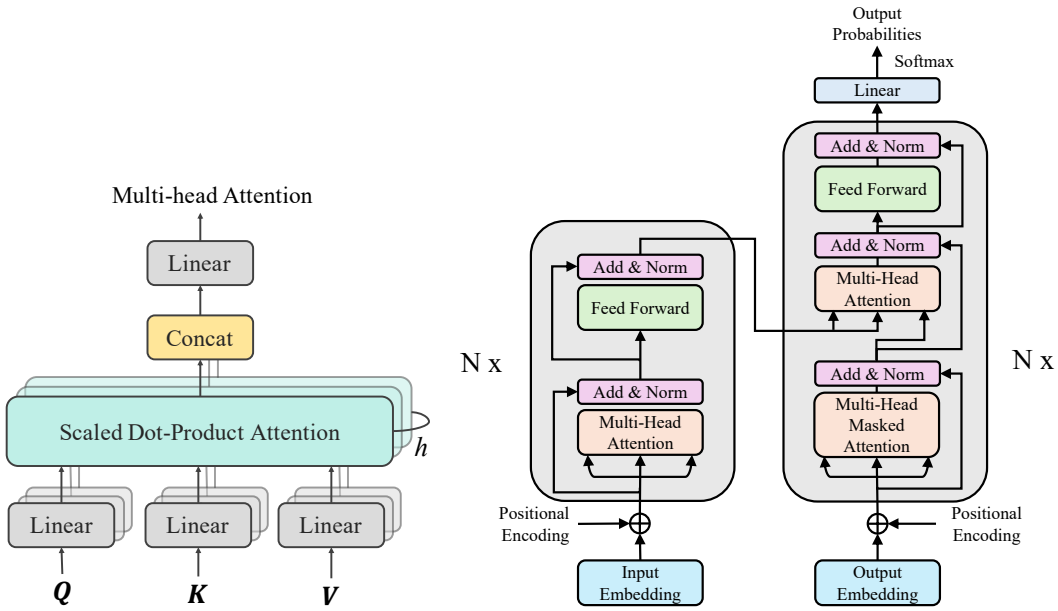
$$\boldsymbol{c}_2 = \text{softmax}(\boldsymbol{\alpha}_2)\boldsymbol{X}\boldsymbol{W}_v, \tag{2.2}$$

where $\text{softmax}(\boldsymbol{\alpha}_2) = \frac{e^{\alpha_{2,i}}}{\sum_{j=1}^{4} e^{\alpha_{2,j}}} (i = 1, \ldots, 4)$ ensures the weights sum to 1. If $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v$ are properly trained, then the importance $\alpha_{2,3}$ corresponding to the word "boy" should be high, which is semantically related to the word "that".

In practice, multiple groups of $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v$ are independently initialized and trained to enrich the vector representation. Each group of $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v$ is called a *head*, and the attention using multiple heads are termed as **multi-head attention**, as in Figure 2.2a.

2.1.2 Architecture of Transformer

Transformer is a multi-layer neural network that features an encoder-decoder structure. It stacks attention and feed-forward layers as illustrated in Figure 2.2b,



(a) Multi-head Attention (b) Model architecture of the transformer

Figure 2.2: Illustrations of Multi-head attention and model architecture of the transformer [93].

where self-attention layers act as the key component. The encoder maps an input sequence $(w_1, \ldots, w_n)$ to an intermediate sequential representation $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, and the decoder generates an output sequence $(y_1, \ldots, y_n)$ one-by-one.

The encoder of a transformer stacks $N = 6$ identical layers, each consisting of two sub-layers. The first sub-layer is an attention layer applying multi-head self-attention on the input, and the second sub-layer is a feed-forward neural network. Residual connection [39] and layer normalization [5] stand between the sub-layers. Specifically, the output of each sub-layer is $\text{LayerNorm}(\boldsymbol{x} + \text{Sublayer}(\boldsymbol{x}))$, where $\boldsymbol{x}$ is the encoder's initial input and $\text{Sublayer}(\cdot)$ is the function inside each sub-layer. Outputs of each sub-layer share the same dimensionality d_m with the input embeddings.

The decoder differs from the encoder in that a third self-attention sub-layer is inserted, attending the contextualized embeddings to the output of the encoder stack. Also, the first self-attention sub-layers are modified to mask those positions after the current position, ensuring the predictions for position i only depend on known outputs before i.

Additionally, position encodings (PE) are added to the input embeddings to incorporate the sequence order. The position encoding is computed using sine and cosine functions:

$$PE_{(pos,2i)} = sin(\frac{pos}{10000^{2i/d_m}}), \tag{2.3}$$

$$PE_{(pos,2i+1)} = cos(\frac{pos}{10000^{2i/d_m}}), \tag{2.4}$$

where pos represents the position index and i represents the dimension index. These functions are chosen because PE_{pos+k} can be expressed as a linear function of PE_{pos}, enabling the model to more easily learn to attend by relative positions [93].

2.2 Language Modeling

2.2.1 Neural Language Model

Language models (LMs) assign probabilities to sequences of words [50]. Given a sequence of N words $(w_1, w_2, \cdots, w_N)$, LMs compute the probability of the sequence $P(w_1, w_2, \cdots, w_N)$. If modeling the probability forwardly, the model is called a forward LM, computing the probability of token w_t given the history $(w_1, \cdots, w_{t-1})$:

$$P(w_1, w_2, \cdots, w_N) = \prod_{t=1}^{N} P(w_t | w_1, w_2, \cdots, w_{t-1}) = \prod_{t=1}^{N} P(w_t | w_{1:t-1}). \tag{2.5}$$

A backward LM runs over the sequence reversely, predicting the previous word given the future context:

$$P(w_1, w_2, \cdots, w_N) = \prod_{t=1}^{N} P(w_t | w_{t+1}, w_{t+2}, \cdots, w_N) = \prod_{t=1}^{N} P(w_{t+1:N}). \tag{2.6}$$

In other words, LMs predict upcoming words from prior word context, following either a forward or a backward direction. During this process, the likelihood of each word given the known words is calculated. Consequently, the likelihood of the entire sequence can be derived from the probabilities of all composing words. LMs modeling from both directions are called **bi**directional **L**anguage **M**odels (**biLM**).

To improve computational efficiency, the likelihood of token w_t is computed by conditioning only a subset of prior words instead of its entire history. This approximation method is called **n-gram**, which looks at $n-1$ words in the past, as shown below.

$$P(w_t|w_{1:t-1}) \approx P(w_t|w_{t-n+1:t-1}). \tag{2.7}$$

Bengio et al. [10] first applied neural network for language modeling. Under their design, feed-forward neural networks are used for constructing an n-gram language model. A feed-forward neural network is the most straightforward multi-layer neural network processing computations from lower layers to higher layers without a loop. In general, each layer of a feed-forward neural network receives an input vector $\boldsymbol{x}$ from the previous layer and generates a vector $\boldsymbol{h}$ as the output:

$$\boldsymbol{h} = \sigma(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}), \tag{2.8}$$

where $\boldsymbol{W}$ is a a projection matrix, $\boldsymbol{b}$ is a bias vector and σ is an activation function with non-linearity.

A feed-forward LM takes input at timestep t, a representation of n previous words, then outputs a probability distribution over possible next words. Mathematically, the function $f(w_t, \cdots, w_{t-n+1}) = P(w_t|w_{1:t-1})$ is decomposed into two parts [10]:

1. A mapping $\boldsymbol{C}$ from any element i of vocabulary $\mathcal{V}$ to a real vector $\boldsymbol{C}(i) \in \mathbb{R}^m$, where m is the dimension of word embeddings. In practice, $\boldsymbol{C}$ is represented by a $|V| \times m$ matrix of trainable parameters.

2. The feed-forward neural network g maps an input sequence of word embeddings $(\boldsymbol{C}(w_{t-n+1}), \cdots, \boldsymbol{C}(w_{t-1}))$ to a conditional probability distribution over words in $\mathcal{V}$ for the next word w_t. The output of g is a vector, with the i-th element yields the probability $P(w_t = i|w_{1:t-1})$.

Combining the two parts above yields:

$$f(i, w_{t-1}, \ldots, w_{t-n+1}) = g(i, \boldsymbol{C}(w_{t-1}), \ldots, \boldsymbol{C}(w_{t-n+1})). \tag{2.9}$$

The model can be trained by minimizing the cross-entropy (negative log-likelihood) loss. At timestep t, suppose the correct next word is $w_t = i$, then:

$$L_{CE} = -\log P(w_t = i|w_{t-1}, ..., w_{t-n+1}). \tag{2.10}$$

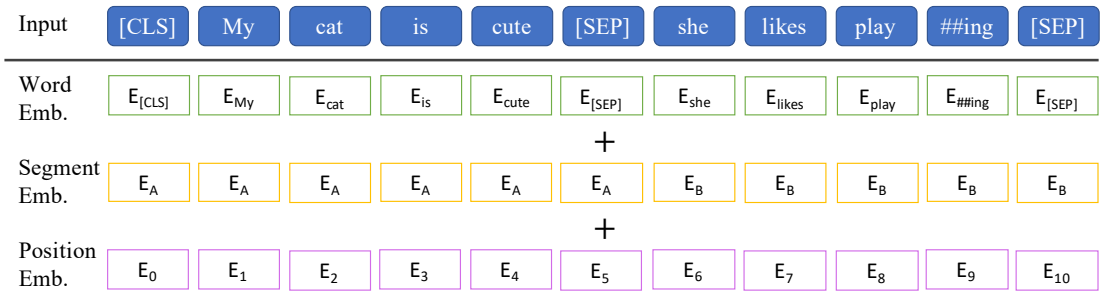


Figure 2.3: Input representation of BERT, where Emb. is short for embedding.

Although modern neural LMs generally utilize architectures other than a feed-forward neural network, feed-forward LMs underlie the design of transformer-based LMs.

2.2.2 Large Language Models (LLMs)

This subsection introduces several pre-trained LLMs closely related to this study, all based on the Transformer architecture described in Section 2.1.

BERT: Bidirectional Encoder Representations from Transformers

Bidirectional **E**ncoder **R**epresentations from **T**ransformers (**BERT**, [28]) is an encoder-only LLM that returns a contextualized representation for each input word. As indicated by its name, the model is a biLM that reads the context from both right to left and left to right. The architecture is a multi-layer bidirectional Transformer encoder, where $\text{BERT}_{\text{BASE}}$ contains 12 Transformer layers with 110M parameters in total, and $\text{BERT}_{\text{LARGE}}$ contains 24 Transformer layers with 340M parameters in total. More details about the architecture and parameters can be found in the original paper [28].

Originally, BERT is pre-trained under two kinds of unsupervised losses:

1. A *masked language model* objective, where a random word in the input is masked, and the goal is to predict the original vocabulary ID of the masked word.

2. A *next sentence prediction* objective, where the goal is to predict if sentence B fluently follows sentence A.

An example input is shown in Figure 2.3. Here, the word "playing" is split into two sub-words, or tokens, as "play" and "##ing". This is because words are broken down into wordpieces to process more words with a limited vocabulary size [102]. For the *masked language model* pre-training objective, 15% of the tokens in each sequence are chosen randomly to be masked with a special token `[MASK]`. For the *next sentence prediction* pre-training objective, 50% of the time, sentence B is the actual sentence following A, and 50% of the time, it is a random sentence from the corpus. Whether B follows A is predicted from the representation corresponding to the beginning `[CLS]` token in the output layer.

Since its release, variants and improvements of BERT have been proposed. For example, Liu et al. [64] have proposed RoBERTa as a robust version of BERT trained with a larger corpus and longer time. Beltagy et al. [8] have proposed SciBERT as a variant of BERT specialized in the science domain, pre-trained on a large multi-domain corpus of scientific publications. Apart from those mentioned above, there are also other works following the framework of BERT [20, 49, 56]. A comprehensive survey of all such works is out of the scope of this study and is not included in this dissertation.

Multilingual BERT. Training BERT on the concatenation of monolingual Wikipedia corpora yields a multilingual version of BERT, shortened as mBERT [28]. mBERT is good at zero-shot cross-lingual model transfer: when fine-tuning the model using task-specific supervised data from one language, it generalizes surprisingly well when evaluated on another language [78]. Conneau and Lample [23] and Conneau et al. [24] incorporated a translation language modeling objective to the masked language model objective, further improving the cross-lingual transferability. The second part of this dissertation adopts mBERT for encoding cross-lingual DocRE.

GPT: Generative Pre-trained Transformer

Generative **P**re-trained **T**ransformer (GPT) is another Transformer-based LLM frequently used for NLP research [80]. GPT is proposed before BERT, whose model architecture is nearly identical to BERT apart from the attention masking. Specifically, although both GPT and $\text{BERT}_{\text{BASE}}$ consist of 12 Transformer layers, GPT comprises Transformer *decoder* blocks, and BERT comprises Transformer

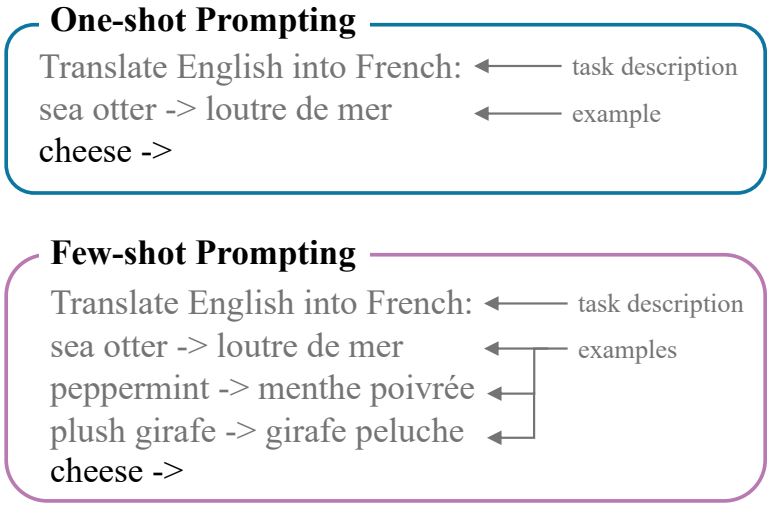


Figure 2.4: Example inputs for in-context learning [13].

encoder blocks. In other words, BERT processes the input from both forwarding and backwarding directions, while GPT processes the input only forwardly, i.e., from left to right. For this reason, BERT is commonly referred to as an encoder, which can be used for encoding texts; GPT is commonly referred to as a decoder, which can be used for generating texts.

GPT has evolved greatly since its proposal. Proceeding GPT is GPT-2 [81], which expanded more than 10 times the parameters of GPT to 1.5B and performed well across multiple domains and datasets. GPT-3 with 175B parameters further replaced GPT-2 with an impressive performance on many NLP tasks, in some cases nearly matching that of state-of-the-art fine-tuned models [13]. They brushed up on the training strategies and published GPT-3.5, which is better at following humans' instructions. Recently, GPT-4 has surprised the world with human-level performance on various professional and academic benchmarks, with the model size not officially announced yet [74]. ChatGPT, the famous chatbot that brings NLP to real-world use, is typically driven by GPT-3.5 or GPT-4.

In-Context Learning of LLMs

As mentioned above, in some cases, GPT-3 can exhibit comparable performance to supervisedly-trained models on several NLP tasks, such as translation, summarization, and question answering. To achieve this, the model learns from few-shot

examples provided by humans to understand the goal of each task. The method is named **I**n-**C**ontext **L**earning (**ICL**,[13]), with an example shown in Figure 2.4. The core idea is an inductive learning process based on analogies [30]. To be more specific, the model reads the task descriptions and few-shot demonstrations provided as the input, which is referred to as the *prompt*, to understand the task and respond to new instances. It has been reported that, with proper descriptions and demonstrations, LLMs can perform many tasks with reasonably high performance [13, 30]. Therefore, ICL is a promising paradigm for controlling the behavior of AI that is handy and training-free.

3 Prelimiaries and Related Work

This chapter delves into the specific task addressed by this study, i.e., Document-level Relation Extraction (DocRE). To this end, Section 3.1 formally defines the task and introduces notations used throughout this dissertation. The section also details the most commonly used dataset and how it is constructed, providing important context for understanding both the model and dataset construction parts. Afterward, related work about the model and dataset construction parts is introduced in Section 3.2 and Section 3.3, respectively.

3.1 Task Definition and Dataset

3.1.1 DocRE: Task Definition

Given a document D containing sentences $\mathcal{X}_D = \{x_1, x_2, \ldots, x_n\} = \{x_i\}_{i=1}^{n}$ and entities $\mathcal{E}_D = \{e_1, e_2, \ldots, e_l\} = \{e_i\}_{i=1}^{l}$, the goal of DocRE is to predict all possible relations between every entity pair. Each entity $e \in \mathcal{E}_D$ is mentioned at least once in D, with all its proper-noun mentions denoted as $\mathcal{M}_e = \{m_1, m_2, \ldots, m_k\} = \{m_i\}_{i=1}^{k}$. An entity pair (e_s, e_o) can hold multiple relations, comprising a set $\mathcal{R}_{s,o} \subset \mathcal{R}$, where $\mathcal{R}$ is a pre-defined relation set. The relation label set $\mathcal{R}$ includes ϵ, which stands for *no-relation*.

Additionally, the task **E**vidence **R**etrieval (**ER**) aims at retrieving the evidence for each relation prediction at sentence level. If an entity pair (e_s, e_o) carries a valid relation $r \in \mathcal{R} \backslash \{\epsilon\}$, ER aims to retrieve the supporting evidence $\mathcal{V}_{s,r,o} \subseteq \mathcal{X}_D$ that are sufficient to predict the triplet (e_s, r, e_o). In general, a model that performs both DocRE and ER should return a quadruple $(e_s, r, e_o, \mathcal{V}_{s,r,o})$ given the document D and an entity pair (e_s, e_o).

	Human-Annotated	Machine-Annotated
# Documents	5,051	101,873
# Relation Types	96	96
# Sentences per Document	8.0	8.1
# Entities per Document	19.5	19.3
# Mentions per Entity	1.3	1.3
# Triples per Document	12.5	14.8
# Evidence per Triple	1.6	–

Table 3.1: Statistics of DocRED collected by Yao et al. [112].

3.1.2 DocRED: Dataset Statistics

Yao et al. [112] first formulated the task DocRE with a large-scale dataset constructed from Wikipedia. The dataset, named DocRED, consists of 5,051 documents with human annotations and 101,873 with automatic annotations, with statistics detailed in Table 3.1. As mentioned in Section 1.2, DocRE is difficult to annotate. Therefore, the study adopted a recommendation-based annotation scheme, where human annotators modified the recommendations provided by the model instead of listing relation triples from scratch[1]. The strategy for recommending relation triples was Distant Supervision [71], which requires aligning text to an existing knowledge base.

Distant Supervision

Distant Supervision is the method that supervises training of relation extraction models with a **K**nowledge **B**ase (**KB**). The method assumes that if a sentence contains an entity pair that participates in a known relation in a knowledge base (KB), the sentence probably expresses that relation. Given a large pool of natural language texts and a pre-defined KB, the method collects training data for relation extraction by aligning the KB to the text [71]. An important assumption here is that if a relation triple (e_s, r, e_o) presents in a pre-defined KB and entities e_s and e_o appear simultaneously in a sentence, then the sentence is likely to express the relation (e_s, r, e_o) and thus included as a training instance for relation r. In such a way, entities are used as anchors to automatically collect

[1]Details of the construction process are described in Section 3.3.

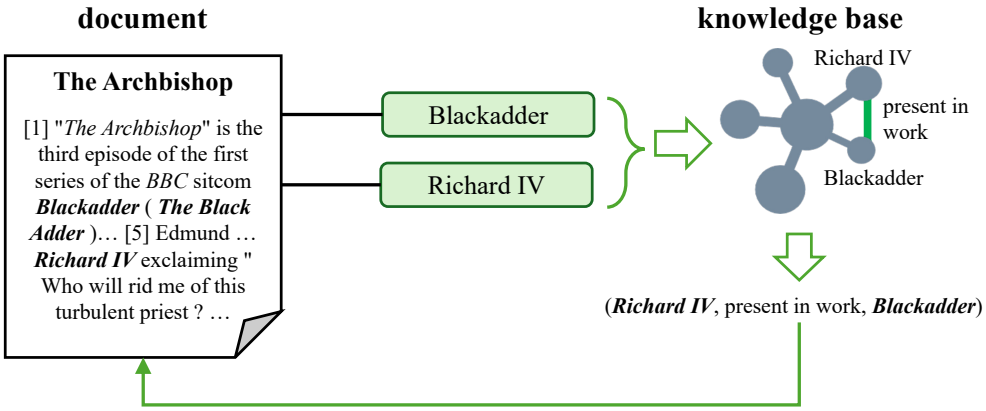


Figure 3.1: Example of obtaining training instances for DocRE via distant supervision.

training instances for RE with no human efforts needed.

Distant Supervision has been widely adopted to generate automatically-labeled data for RE [71, 79, 103]. Figure 3.1 showcases how Yao et al. [112] collected automatically-annotated data with distant supervision. Specifically, they aligned Wikidata [96], a KB based on Wikipedia pages maintained by humans, with introductory sections from Wikipedia pages. Named Entity Recognition (NER) was first conducted to identify all entities within each document. Then, entity pairs were enumerated as queries for the KB to see if an edge connected the entities. If the entity pair were connected in the KB, the triple would be included as an automatically annotated relation instance in the given document.

3.2 Model Construction

This section introduces several representative DocRE models related to this study. All these models are based on Transformers as introduced in Section 2.1. These transformer-based models outperform their graph-based counterparts [116, 117, 107], possibly owing to the capability of Transformer encoders in capturing long-distance token (word) dependencies.

Specifically, three models are introduced in the following section: ATLOP [124], EIDER [105], and KD-DocRE [88]. ATLOP is the backbone model of the proposed method; EIDER is an important competitor with the proposed method that

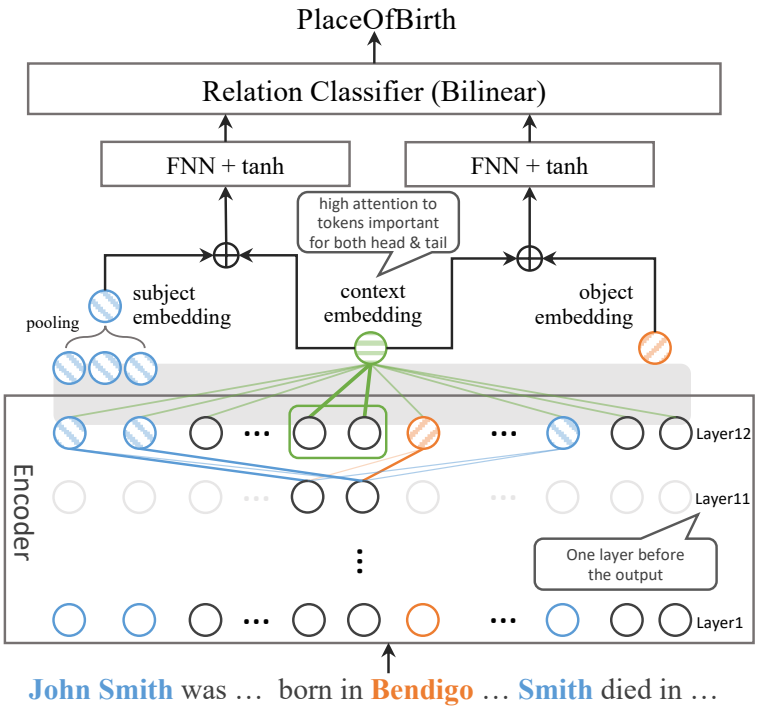


Figure 3.2: Model architecture of ATLOP [124]. FNN is short for Feed-forward Neural Network.

tackles both DocRE and ER; and KD-DocRE was the state-of-the-art DocRE model before the proposed method, which utilizes both human-annotated and machine-annotated parts of DocRED.

3.2.1 ATLOP

ATLOP [124] is an effective and efficient DocRE model based on BERT [28], acting as the backbone of the proposed method. The overall structure is shown in Figure 3.2.

Text Encoding Before encoding, a special token $*$ is inserted at the beginning and the end of each entity mention. Then, tokens $\mathcal{T}_D = \{t_i\}_{i=1}^{|\mathcal{T}_D|}$ within document D are encoded with a Transformer-based pretrained language model (PLM, [93]) to obtain token embeddings and cross-token dependencies. Notably, although the original ATLOP adopts only the last layer, this work takes the average of the last

three layers, as pilot experiments showed that using the last 3 layers yields better performance than using only the last layer. Specifically, for a PLM with d hidden dimensions at each transformer layer, the token embeddings $\boldsymbol{H}$ and cross-token dependencies $\boldsymbol{A}$ are computed as:

$$\boldsymbol{H}, \boldsymbol{A} = \text{PLM}(\mathcal{T}_D), \tag{3.1}$$

where $\boldsymbol{H} \in \mathbb{R}^{|\mathcal{T}_D| \times d}$ averages over hidden states of each token from the last three layers and $\boldsymbol{A} \in \mathbb{R}^{|\mathcal{T}_D| \times |\mathcal{T}_D|}$ averages over attention weights of all attention heads from the last three layers.

Entity Embedding The entity embedding $\boldsymbol{h}_e \in \mathbb{R}^d$ for each entity e with mentions $\mathcal{M}_e = \{m_i\}_{i=1}^{|\mathcal{M}_e|}$ is computed by collecting information from all its mentions. Specifically, `LogSumExp` pooling, which has been empirically shown to be effective in previous studies [48], is adopted as:

$$\boldsymbol{h}_e = \log \sum_{i=1}^{|\mathcal{M}_e|} \exp(\boldsymbol{H}_{m_i}), \tag{3.2}$$

where $\boldsymbol{H}_{m_i}$ is the embedding of the special token $*$ at the starting position of mention m_i.

Localized Context Embedding To better utilize information from long texts, ATLOP introduces entity-pair specified localized context embeddings. Intuitively, for entity pair (e_s, e_o), tokens important to both e_s and e_o should contribute more to the embedding. The importance of each token is determined by the cross-token dependencies $\boldsymbol{A}$ obtained from Equation 3.1. For entity e_s, the importance of each token is computed using the cross-token dependencies of all its mentions $\mathcal{M}_{e_s}$. First, ATLOP collects and averages over the attention $\boldsymbol{A}_{m_i} \in \mathbb{R}^{|\mathcal{T}_D|}$ at the special token $*$ before each mention $m_i \in \mathcal{M}_{e_s}$ to get $\boldsymbol{a}_s \in \mathbb{R}^{|\mathcal{T}_D|}$ as the importance of each token for entity e_s. Then, the importance of each token for an entity pair (e_s, e_o), noted as $\boldsymbol{q}^{(s,o)} \in \mathbb{R}^{|\mathcal{T}_D|}$, is computed from $\boldsymbol{a}_s$ and $\boldsymbol{a}_o$ as:

$$\boldsymbol{q}^{(s,o)} = \frac{\boldsymbol{a}_s \circ \boldsymbol{a}_o}{\boldsymbol{a}_s^\top \boldsymbol{a}_o}, \tag{3.3}$$

where $\circ$ stands for the Hadamard product. $\boldsymbol{q}^{(s,o)}$ is thus a distribution that reveals the importance of each token for entity pair (e_s, e_o). Subsequently, ATLOP

performs a localized context pooling,

$$\boldsymbol{c}^{(s,o)} = \boldsymbol{H}^\top \boldsymbol{q}^{(s,o)}, \tag{3.4}$$

where $\boldsymbol{c}^{(s,o)} \in \mathbb{R}^d$ is a weighted average over all token embeddings.

Relation Classification To predict the relation between entity pair (e_s, e_o), ATLOP first generates context-aware subject and object representations:

$$\boldsymbol{z}_s = \tanh(\boldsymbol{W}_s[\boldsymbol{h}_s; \boldsymbol{c}^{(s,o)}] + \boldsymbol{b}_s) \tag{3.5}$$

$$\boldsymbol{z}_o = \tanh(\boldsymbol{W}_o[\boldsymbol{h}_o; \boldsymbol{c}^{(s,o)}] + \boldsymbol{b}_o), \tag{3.6}$$

where $[\cdot;\cdot]$ represents the concatenation of two vectors and $\boldsymbol{W}_s, \boldsymbol{W}_o \in \mathbb{R}^{d\times 2d}, \boldsymbol{b}_s, \boldsymbol{b}_o \in \mathbb{R}^d$ are trainable parameters. Then, a bilinear classifier[2] is applied on the context-aware representations to compute the relation scores $\boldsymbol{y}_{s,o} \in \mathbb{R}^{|\mathcal{R}|}$:

$$\boldsymbol{y}_{s,o} = \boldsymbol{z}_s^\top \mathsf{W}_r \boldsymbol{z}_o + \boldsymbol{b}_r, \tag{3.7}$$

where $\mathsf{W}_r \in \mathbb{R}^{|\mathcal{R}|\times d\times d}$ and $\boldsymbol{b}_r \in \mathbb{R}^{|\mathcal{R}|}$ are trainable parameters. The probability that relation $r \in \mathcal{R}$ holds between entity e_s and e_o is thus computed from:

$$\mathrm{P}(r|s,o) = \mathrm{sigmoid}(y_{s,r,o}). \tag{3.8}$$

Loss Function ATLOP proposes Adaptive Thresholding Loss (ATL), which learns a dummy threshold class TH during training, serving as a dynamic threshold for each relation class $r \in \mathcal{R}$. For each entity pair (e_s, e_o), ATL forces the model to yield scores above TH for positive relation classes $\mathcal{R}_P$ and scores below TH for negative relation classes $\mathcal{R}_N$, formulated as below:

$$\begin{aligned}\mathcal{L}_{\mathrm{RE}} = &-\sum_{s\neq o}\sum_{r\in\mathcal{R}_P} \frac{\exp(y_{s,r,o})}{\sum_{r'\in\mathcal{R}_P\cup\{\mathrm{TH}\}}\exp(y_{s,r',o})} \\ &- \frac{\exp(y_{s,\mathrm{TH},o})}{\sum_{r'\in\mathcal{R}_N\cup\{\mathrm{TH}\}}\exp(y_{s,r',o})}.\end{aligned} \tag{3.9}$$

The idea of setting a threshold class is similar to the Flexible Threshold [15]. During inference, relation classes with scores higher than the threshold class are extracted as relation predictions for each entity pair.

[2]In practice, a grouped bilinear classifier [121] is applied to save memory.

3.2.2 EIDER

EIDER extends ATLOP with an Evidence Extraction module [105]. While ATLOP deals with only DocRE, EIDER incorporates an *evidence classifier* so that the model tackles both DocRE and ER. Notably, EIDER identifies the evidence *entity-pair-wise* instead of triple-wise. In other words, for an entity pair (e_s, e_o) holding relations r_1, r_2, the model extracts only one set of evidence $\mathcal{V}_{s,o}$ instead of two sets of evidence $\mathcal{V}_{s,r_1,o}$ and $\mathcal{V}_{s,r_2,o}$. Such a design was chosen because the authors observed that most entity pairs have only one set of evidence across relations.

The following subsection details the incremental elements of EIDER compared to ATLOP, namely the **evidence classifier** to extract the evidence sentences and the **inference-stage fusion** strategy to fuse predicted results from the whole document and a partial document composing only evidence sentences.

Evidence Classifier The model specifies the evidence of each entity pair via a bilinear classifier, similar to that of the relation classifier. To be specific, they first obtain an embedding $\boldsymbol{x}_i \in \mathbb{R}^d$ for each sentence x_i by applying a `LogSumExp` pooling over all composing tokens $\mathcal{T}_{x_i}$:

$$\boldsymbol{x}_i = \log \sum_{i=1}^{|\mathcal{T}_{x_i}|} \exp(\boldsymbol{H}_{t_i}), \tag{3.10}$$

where $\boldsymbol{H}_{t_i}$ is the embedding of the t_i-th token that composes x_i.

Then, the sentence embedding $\boldsymbol{x}_i$, together with the localized context embedding $\boldsymbol{c}^{(s,o)}$ concerning e_s, e_o, are fed into an evidence classifier to decide the likelihood of sentence x_i being the evidence of relation decisions of entity pair e_s, e_o:

$$\mathrm{P}(x_i | e_s, e_o) = \mathrm{sigmoid}(\boldsymbol{x}_i^\top \boldsymbol{W}_v \boldsymbol{c}^{(s,o)} + b) \tag{3.11}$$

where $\boldsymbol{W}_v \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}$ are trainable parameters. The evidence classifier is thus trained using the binary cross-entropy loss as below.

$$\mathcal{L}_{\mathrm{ER}} = -\sum_{s \neq o, r \neq \epsilon} \sum_{x_i \in \mathcal{X}_D} v_i^{(s,r,o)} \cdot \mathrm{P}(x_i | e_s, e_o) + (1 - v_i^{(s,r,o)}) \cdot \log(1 - \mathrm{P}(x_i | e_s, e_o)), \tag{3.12}$$

where $v_i^{s,r,o} = 1$ if sentence x_i is an evidence of relation triple (e_s, r, e_o), i.e., $x_i \in \mathcal{V}_{s,r,o}$, otherwise $v_i^{s,r,o} = 0$.

The overall loss function of EIDER is a weighted loss of $\mathcal{L}_{\text{RE}}$ and $\mathcal{L}_{\text{ER}}$:

$$\mathcal{L} = \mathcal{L}_{\text{RE}} + \lambda \mathcal{L}_{\text{ER}}, \tag{3.13}$$

where λ is empirically set to 0.1.

Having introduced the mechanism of evidence classifier, it is clearer why EIDER marginalizes relation labels for each entity pair during ER. Under their design, it is necessary to learn $|\mathcal{R}|$ representations for each sentence as in Equation 3.10 if considering each relation label separately. This results in expensive computation, given $|\mathcal{R}| = 96$ in Table 3.1.

Inference-Stage Fusion EIDER proposes fuse relation prediction results from (1) the original whole document D and (2) a group of partial documents obtained by collecting the predicted evidence sentences. The motivation is to utilize both the original document and predicted results from the evidence classifier. While an ideal evidence classifier should be able to pick up exact evidence sentences sufficient for deciding the relation of each entity pair, obtaining such a classifier in real-world use is impractical. Therefore, predictions from the whole document are also included to compensate for potential information loss from an un-ideal evidence classifier.

Details of inference-stage fusion are summarized in Figure 3.3. During inference, firstly, the model receives the whole document D as the input and returns the relation extraction and evidence retrieval results $(e_s, \hat{r}, e_o, \hat{\mathcal{V}}_{s,\hat{r},o})$ for each entity pair (e_s, e_o). Logit $y^D_{s,\hat{r},o}$ representing the likelihood of relation $(e_s, \hat{r}, e_o)$ presents in document D is obtained from Equation 3.8.

Next, for each predicted quadruplet $(e_s, \hat{r}, e_o, \hat{\mathcal{V}}_{s,\hat{r},o}) \in \mathcal{P}^D$, a partial document $\hat{D}$ is constructed by concatenating $\hat{\mathcal{V}}_{s,\hat{r},o}$. $\hat{D}$ is expected to contain only sentences useful for extracting the relation $(e_s, \hat{r}, e_o)$. The operation theoretically results in $|\mathcal{P}^D|$ partial documents, while some may be duplicated to another, i.e., containing the same set of sentences. Each partial document $\hat{D}$ is then fed into the model to obtain relation extraction results, from which the logit $y^{\hat{D}}_{s,\hat{r},o}$ for each triple $(e_s, \hat{r}, e_o)$ can be obtained.

To finally decide if $(e_s, \hat{r}, e_o)$ is a valid relation triple, the logit is computed from summing up all logits regarding $(e_s, \hat{r}, e_o)$, both in the original document D

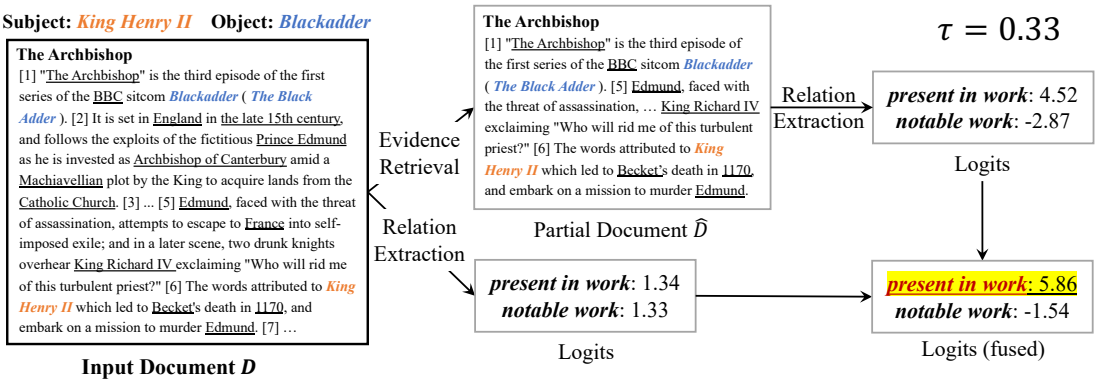


Figure 3.3: An example of Inference-Stage Fushion. The fused logits are the summation of corresponding logits of the input and the partial documents. The final prediction will be (King Henry II, *present in work*, Blackadder).

and in all partial documents:

$$y_{s,\hat{r},o} = \sum_{D' \in D \cup \Delta} y^{D'}_{s,\hat{r},o}, \tag{3.14}$$

where $\Delta = \{\hat{D}_1, \hat{D}_2, \ldots, \hat{D}_p\}$ are partial documents. Relation triple $(e_s, \hat{r}, e_o)$ is considered true only if the summation in Equation 3.14 surpasses the threshold τ. Here, τ is not a fixed pre-defined hyper-parameter but a parameter chosen to maximize the relation extraction F1 score on the development set, optimized separately for each model.

Other Studies about ER in DocRE Apart from EIDER, several studies have also tackled ER in DocRE. For example, Huang et al. [43] first reported that heuristically selecting evidence sentences boosts the performance of DocRE models. E2GRE [41] and SAIS [104] incorporate neural classifiers to retrieve evidence together with RE automatically. Unlike EIDER, these models retrieve evidence triple-wise, i.e., specify evidence sentences regarding each relation label. As a result, the computation costs of training E2GRE and SAIS are higher than that of EIDER.

3.2.3 KD-DocRE

KD-DocRE [88] is another DocRE model based on ATLOP, which scored highest among all DocRE models before DREEAM was published. The method addresses only DocRE without considering the evidence through a three-fold proposal: (1) A method for representation learning, (2) a loss function for training DocRE models, and (3) a strategy for utilizing distantly supervised data. This subsection details the third one, i.e., utilizing the distantly supervised data provided in DocRED with a knowledge distillation scheme, which is the most relevant to this study.

Knowledge Distillation KD-DocRE distills knowledge from a model trained on human-annotated data while utilizing the distantly supervised data for training. To do so, the authors introduced two models with the same configuration, one noted as the teacher model and the other as the student model. The teacher model is trained on the human-annotated data, which generates soft relation labels on distantly-supervised data. Then, the student model learns the soft relation labels and the automatically annotated labels via distant supervision.

Specifically, the objective function is to minimize the **M**ean **S**quared **E**rror (**MSE**) between the soft labels generated from the teacher model and the predicted logits from the student model. For each entity pair (e_s, e_o), the loss of knowledge distillation is defined as:

$$\mathcal{L}_{\text{KD}} = MSE(\boldsymbol{y}^S_{s,o}, \boldsymbol{y}^T_{s,o}), \tag{3.15}$$

where $\boldsymbol{y}^S_{s,o}$ is the logits predicted from the student model and $\boldsymbol{y}^T_{s,o}$ is the soft labels generated from the teacher model.

The overall loss of the student model trained on the distantly supervised data is thus the summation of knowledge distillation and relation extraction losses:

$$\mathcal{L}^S = \mathcal{L}_{\text{KD}} + \mathcal{L}_{\text{RE}}, \tag{3.16}$$

where $\mathcal{L}_{\text{RE}}$ represents the relation extraction loss using the automatically annotated labels as the ground truth. According to the original paper, such a training strategy outperforms that of using only $\mathcal{L}_{\text{RE}}$ to train a model on the distantly-supervised data [88].

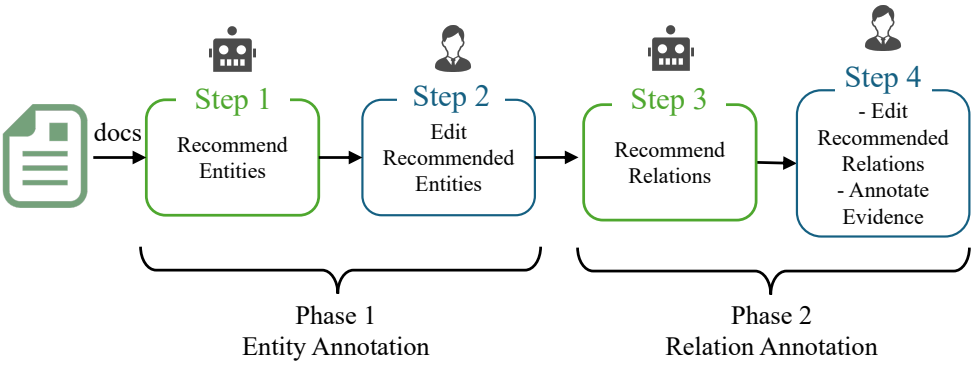


Figure 3.4: Annotation pipeline described in DocRED [112].

3.3 Dataset Construction

This section reviews several studies related to the second part of the dissertation, i.e., reducing the human annotation costs of constructing a DocRE dataset with the assistance of existing datasets in another language. To this end, the section discusses existing DocRE datasets and their construction process. The datasets are categorized by language, with English datasets introduced first, followed by non-English ones. Next, the section presents existing works that use cross-lingual techniques for structured prediction tasks. The automatically annotated dataset mentioned in this study is constructed using a similar strategy.

3.3.1 DocRE Datasets in English

The definition of general-purpose DocRE was proposed by Yao et al. [112], along with DocRED, a dataset constructed from the English Wikipedia. While two document-level relation extraction datasets, namely CDR [59] and GDA [101], have been proposed ahead of DocRED, they were collected in the biomedical domain, thus unsuitable for developing general-purpose DocRE models.

Pipeline of Collecting DocRED Constructing DocRED involves two phases: the entity annotation phase and the relation annotation phase. An overview is shown in Figure 3.4. The entity annotation phase contains the following steps:

1. A pre-trained Named Entity Recognition (NER) model[3] automatically de-

[3] `https://spacy.io`

tected entity mentions in each document.

2. Human annotators reviewed the recognized entity spans, corrected the wrong ones, and supplemented the missed ones. The annotators also merged entity mentions referring to the same entity into one to conduct coreference resolution, deciding the entity mention set $\mathcal{M}_e$ for each entity e.

Then, the relation annotation phase was conducted with the following steps:

1. Automatic annotation based on distant supervision. Specifically, each entity recognized in the entity annotation phase was linked to an item in Wikidata. Then, the relation label(s) on the edge(s) connecting each entity pair were provided as relation recommendations. They also supplemented several recommendations from RE models, but the details are not mentioned.

2. Human annotators reviewed the recommended relation triples, corrected the wrong ones, and supplemented the missed ones. The annotators also picked up all sentences that support the reserved relation instances as evidence.

Limitations of DocRED A critical limitation of DocRED is that it suffers from the **false negative issue**. False negative is a term used to describe when an instance should be classified as positive but is classified as negative. In the case of DocRED, the false negative issue means that a considerable amount of relation instances are absent from the ground-truth annotations [44, 89, 105].

The issue of incomplete relation annotations comes from the complexity of DocRE. A document with an average length of 200 tokens containing multiple sentences is semantically more complex than a single sentence. The average number of entities per document is 19.5 (ref. Table 3.1), which may comprise a quadratic number of entity pairs, i.e., $19.5 \times 19.5 \approx 380$, potentially carrying relation(s). While manually listing up all relation instances from scratch is infeasible, the recommendation-based annotation strategy puts annotators into a dilemma: Compared with supplementing missing relation triples, validating the correctness of existing triples is much easier. As a result, although instructed to supplement missing triples, annotators tend to perform the easier goal of editing existing ones [44].

Improvements over DocRED Efforts have been paid to alleviate the false negative issue of DocRED. Typically, Huang et al. [44] randomly selected 96 documents from DocRED and relabeled them from scratch, and Tan et al. [89] revised the whole dataset as Re-DocRED with machine assistance.

Huang et al. [44] showcased that half of the relation triples remain missing even with the recommend-revise process conducted in collecting DocRED (Figure 3.5). The missing triples were not included in the machine recommendation, possibly due to the low frequency in Wikidata. Supporting evidence is that for symmetric relations "present in work" and "characters", the latter was missed more frequently than the former. Correspondingly, "present in work" has 165,751 statements, more than the number of "characters" (147,765 statements) in Wikidata[4]. Neither did human annotators supplement the missing triples, indicating the dilemma of the annotation scheme used in DocRED. The observation demonstrated the limitation of automatic annotation with distant supervision.

Tan et al. [89] proposed an iterative approach to alleviating the false negative issue of DocRED. In general, their approach consists of two steps: (1) Automatically generating more relation triples using trained DocRE models and (2) Manually verifying if each generated triple is correct. To improve the diversity of recommendations, they trained three different DocRE models and merged predictions of all models together as candidates for human verification. Relation triples already included in DocRED were treated as correct and excluded from the re-annotation process. Such a process yields Re-DocRED, a re-annotated version of DocRED containing more than twice the relation triples, as in Table 1.1.

3.3.2 DocRE corpora in other languages

DocRE datasets have also been constructed in Chinese and Korean. Cheng et al. [18] constructed HacRED from Chinese DBpedia to promote relation extraction from complex contexts. Yang et al. [111] focused on Korean historical RE research and constructed HistRED from a travel diary written between the 16th and 19th centuries. These datasets, with statistics shown in Table 1.1, were constructed independently from (Re-)DocRED with distinct domains and label sets. This, as introduced in the previous section, results in high annotation costs and difficulty in controlling the quality of constructed datasets.

[4]Source: `https://prop-explorer.toolforge.org/`.

Apart from these studies, Cheng et al. [17] released a system for medical relation extraction on Japanese documents, while the dataset is not publicly available.

3.3.3 Cross-Lingual Projection

The disparity in language resources between English and non-English languages has been a long-standing issue. One of the major techniques to alleviate the disparity is machine translation, where data can be synthesized from one language to another with no human efforts [12, 31, 32]. This section introduces the strategy used for tasks that involve span-level annotations, e.g., part-of-speech tagging, semantic role labeling, or information extraction. For these tasks, merely translating the context from one language to another does not conclude the synthesizing process. An extra step needs to be conducted to project the annotation span accordingly.

Notably, there exist two directions for annotation projection. One is to project from a high-resource language (e.g., English) to a low-resource language (e.g., Arabic), with the purpose of training models that solve tasks in the low-resource language. The other is to project from a low-resource language to a high-resource language, with the purpose of solving the task with models trained in the high-resource language. Both directions are feasible utilizing the same projection methodology.

Align-Based Projection. A common approach for the annotation projection is based on word alignment [2, 52, 69, 73, 113]. Given a dataset in the source language *src* with texts and span-level annotations, the flow of obtaining a parallel dataset in the target language *tgt* goes as follows:

1. Translate texts in *src* into *tgt* with a machine translator;
2. For each sentence pair in *src* and *tgt*, run a word alignment tool to obtain the word-to-word alignment;
3. Based on the word-level alignment, apply heuristics to map the span annotations from *src* to *tgt*.

Such a pipeline introduces two artifacts: a machine translator and a word aligner. The method is thus highly sensitive to error propagation [2]. Furthermore, the

heuristic of mapping the span annotation is non-trivial, as the alignments tend to be non-contiguous, especially for distant language pairs [32, 118].

Mark-Based Projection. Recently, efforts have been made for alignment-free annotation projections [16, 62, 123]. These studies remove the word alignment from the pipeline by using placeholders. Liu et al. [62] and Zhou et al. [123] conducted translation twice to perform cross-lingual Named Entity Recognition with the following approach.

Consider a scenario where the target is to conduct NER in German. The sentence is translated into English, on which English NER models can be applied. The following showcases the alignment-free translation process, where the entity span is marked in blue:

- Bruce Willis wurde in Westdeutschland geboren.

The pipeline of mark-based projection is:

1. Replace span with placeholder and translate:
   Bruce Willis wurde in SPAN geboren. ⇒ Bruce Willis was born in SPAN.

2. Translate the span:
   Westdeutschland ⇒ West German

3. Substitute the placeholder back with the translated span:
   Bruce Willis was born in West German.

After the projection, the probability distribution of the label of Westdeutschland is computed using that of West German.

Chen et al. [16] further merged these two rounds of translations into one with a mark-then-translate strategy. The method surrounds the entity span with its entity label and translates the sentence into another language with the label preserved. For the same example, the original sentence is reformed as:

- Bruce Willis wurde in <LOC> Westdeutschland <\LOC> geboren.

where <LOC> represents the entity label LOCATION. A machine translator was fine-tuned to translate the sentence directly into the target language, e.g., English, as:

- Bruce Willis was born in <LOC> West German <\LOC>.

The study explored several markers, e.g., <>, [], {}, etc., concluding that XML tags work the best with little language-specific semantic meanings.

MultiTACRED Hennig et al. [40], a multilingual version of TACRED [120], was constructed following the same idea. They have efficiently extended the dataset into 12 languages and confirmed its quality to be high enough even without human modifications.

[1] Michael Imperioli[0] (born March 26 , 1966[1]) is an American[2] actor , writer and director best known for his role as Christopher Moltisanti[3] on The Sopranos[4], for which he won the Primetime Emmy Award for Outstanding Supporting Actor[5] in a Drama Series in 2004[6].
[2] He also appeared in the TV drama series Law & Order[7] as NYPD[8] Detective Nick Falco[9].
[3] Imperioli[0] spent the 2008[10] – 2009[11] television season as Detective Ray Carling[12] in the US[13] version of Life on Mars[14].
[4] He was starring as Detective Louis Fitch[15] in the ABC[16] police drama Detroit 1-8-7[17] until its cancellation.
[5] He wrote and directed his first feature film , The Hungry Ghosts[18], in 2008[10].
[6] In 2015[19], he starred in Mad Dogs[20], a dark-comic thriller television series available for viewing on Amazon[21]'s Amazon Prime subscription service[22] in the U.S[13] and on Shomi[24] in Canada[25].

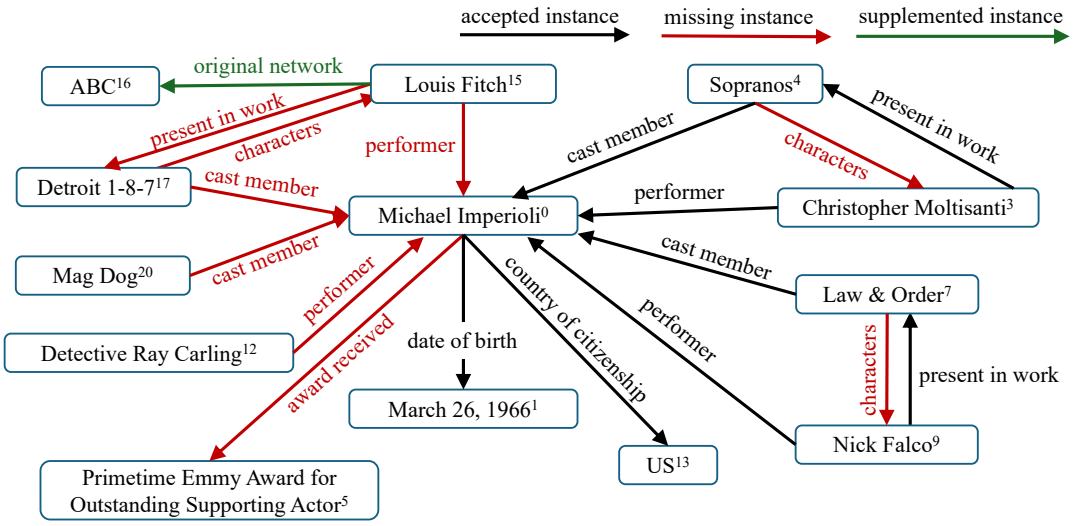


Figure 3.5: A case study of how annotations in DocRED are revised by Huang et al. [44]. The upper part shows the original document, and the lower part shows the annotated relation triples related to entity *Michael Imperioli*. The colors of entities represent their types (PER, TIME, ORG, LOC, MISC). Instances in DocRED rejected by annotators are not shown in this figure.

4 Model Construction: DREEAM

Document-level relation extraction (DocRE) has been recognized as a more realistic and challenging task compared with its sentence-level counterpart [77, 94, 112]. In DocRE, an entity can have multiple mentions scattered throughout a document, and relationships can exist between entities in different sentences. DocRE models are expected to apply information filtering to long texts by focusing more on sentences relevant to the current decision of relation extraction (RE) and less on irrelevant ones.

To this end, existing studies retrieve evidence, a set of sentences necessary for humans to identify the relation between an entity pair [41, 43, 104, 105, 108]. The task of retrieving evidence for relation extraction decisions is named **E**vidence **R**etrieval (ER). Previous studies tackle ER and DocRE as separate tasks, introducing extra neural network layers as evidence classifiers to learn ER. This results in two limitations: (1) task dependencies between ER and DocRE cannot be considered, and (2) high computation cost in training an evidence classifier. As introduced in Section 3.2.2, the evidence classifier is typically a bilinear classifier that receives entity-pair-specific embeddings and sentence embeddings as the input. To compute the evidence score of each sentence for each entity pair, the classifier must walk through all (entity pair, sentence) combinations. The computations significantly increase memory consumption, particularly in documents with numerous sentences and entities. Additionally, the availability of human annotations of evidence is limited. Although silver training data for RE can be automatically collected by distant supervision [71, 112], locating evidence for a sliver RE instance in the document is non-trivial.

This study aims at alleviating these issues to improve the usage of ER in DocRE. To better model the dependency between relation extraction and evidence retrieval, this study proposes **D**ocument-level **R**elation **E**xtraction with

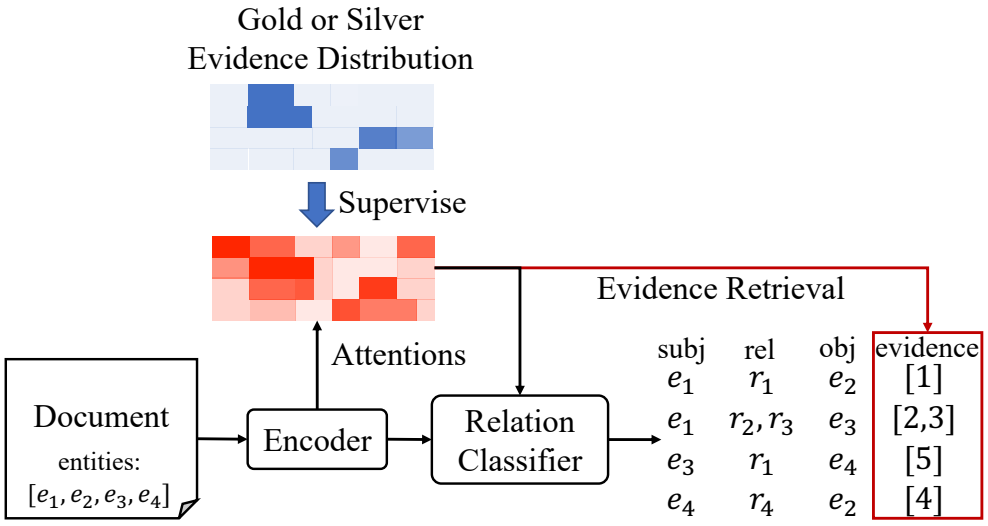


Figure 4.1: Model architecture of DREEAM. Gold/silver evidence distributions come from human-annotations/the teacher model.

Evidence-guided **A**ttention **M**echanism (DREEAM), a memory-efficient approach for incorporating DocRE with ER. DREEAM adopts the basic structure of ATLOP (Section 3.2.1,[124]), a Transformer-based DocRE system widely used in previous studies [88, 104, 105], as the backbone. An overview of DREEAM is shown in Figure 4.1. Instead of introducing an external evidence classifier, DREEAM directly guides the DocRE model in focusing on evidence. To this end, the computation of entity-pair-specific local context embeddings is directly supervised by evidence annotations. The local context embedding, formed as a weighted sum among all token embeddings based on attention from the encoder (Section 3.2.1), is trained to assign higher weights to evidence and lower weights otherwise.

The proposed method can also compensate for the shortage of evidence annotations. Specifically, DREEAM can assign sub-optimal, i.e., silver, evidence on massive, unlabeled data, enabling the weakly-supervised training of ER. The data is obtained from distant supervision (machine-annotated data hereafter) and thus is automatically annotated with relation labels but not evidence labels. This study proposes automatically assigning evidence labels to the machine-annotated data. To this end, a teacher model trained on human-annotated data is adopted to retrieve silver evidence from machine-annotated data. Next, a student model

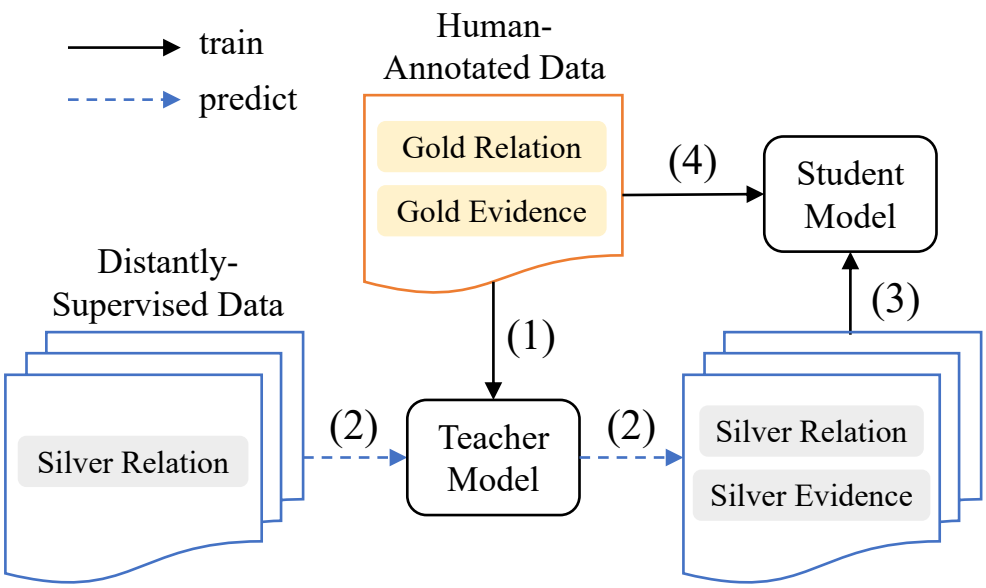


Figure 4.2: Information flow of weakly-supervised training of both DocRE and ER using DREEAM. Arrows represent the direction of knowledge transfer.

is trained on the data for RE while learning ER from the silver evidence. The student model is further finetuned on the human-annotated data to refine its knowledge. Notably, the student model is expected to eventually outperform the teacher model in both DocRE and ER. Experiments on the DocRED benchmark [112] show that with the help of ER weakly-supervised training, DREEAM exhibits state-of-the-art performance on both DocRE and ER. The overview of the proposed weakly-supervised strategy is shown in Figure 4.2, where steps (1)(2)(3)(4) are detailed in Section 4.1.

In short, the contributions of the model construction part are:

- The proposal of DREEAM, a memory-efficient approach to incorporate evidence information directly into Transformer-based DocRE models by guiding the attention. This enables updating the parameters of a DocRE model using the evidence information.
- The state-of-the-art performance of DREEAM on both DocRE and ER. Notably, the method is more memory-efficient than existing approaches, as it eliminates the need for training and inference of the evidence classifier.

- The proposal incorporates distantly supervised DocRE training with weakly supervised ER training, enhancing performance in both tasks. This study is the first to conduct joint training of DocRE and ER under a weakly supervised setting.

This chapter is organized as follows. Section 4.1 presents the details of the proposed method, Section 4.2 reports the experiment settings and results, and Section 4.3 provides detailed analyses about various aspects of DREEAM, the proposed method.

4.1 Proposed Method

ATLOP computes a localized context embedding based on attention weights from the Transformer-based encoder to perform information filtering. The rationale is that cross-token dependencies are encoded as attention weights in Transformer layers. This study proposes DREEAM to enhance ATLOP with evidence: Besides the automatically-learned cross-token dependencies, the attention modules are supervised to concentrate more on evidence sentences and less on others. In this way, the learned DocRE model can decide relation labels based on the evidence information.

DREEAM can be employed for both supervised and weakly-supervised training, sharing the same architecture with different supervisory signals, as shown in Figure 4.1. The pipeline for weakly-supervised training is shown in Figure 4.2, consisted of the following steps:

1. Train a teacher model on human-annotated data with gold relations and evidence labels.

2. Apply the trained teacher model to predict silver evidence for the machine-annotated data.

3. Train a student model on the distantly supervised data, with ER supervised by the silver evidence.

4. Finetune the student model on the human-annotated data to refine its knowledge.

The rest of this section introduces the architecture of DREEAM and how to utilize the model for weakly-supervised training, followed by the inference process.

4.1.1 DREEAM

For each entity pair (e_s, e_o), this study guides the importance of each token $\boldsymbol{q}^{(s,o)} \in \mathcal{R}^{|\mathcal{T}_D|}$ (Equation 3.3) with an **evidence distribution** to help generate an evidence-centered localized context embedding. While $\boldsymbol{q}^{(s,o)}$ yields token-level importance for e_s and e_o, only sentence-level evidence from human annotations is available in DocRED. To alleviate this gap, the importance of each sentence is computed as a sum of the weight of all its composing tokens. Specifically, for a sentence $x_i \in \mathcal{X}_D$ consisting of tokens $t_{\text{START}(x_i)}, \ldots, t_{\text{END}(x_i)}$, the sentence-level importance is computed as:

$$p_i^{(s,o)} = \sum_{j=\text{START}(x_i)}^{\text{END}(x_i)} q_j^{(s,o)}. \tag{4.1}$$

Collecting the importance of all sentences yields a distribution $\boldsymbol{p}^{(s,o)} \in \mathbb{R}^{|\mathcal{X}_D|}$ that expresses the importance of each sentence within the document.

$\boldsymbol{p}^{(s,o)}$ for each entity pair (e_s, e_o) is supervised using a human-annotated evidence distribution computed from gold evidence. First, a binary vector $\boldsymbol{v}^{(s,r,o)} \in \mathbb{R}^{|\mathcal{X}_D|}$ is defined for each valid relation label $r \in \mathcal{R}_{s,o} \subset \mathcal{R} \backslash \{\epsilon\}$ that records whether each sentence $x_i \in \mathcal{X}_D$ is evidence of the relation triple (e_s, r, e_o) or not. For example, if x_i is evidence of (e_s, r, e_o), then $v_i^{(s,r,o)}$ is set to 1, and otherwise 0.

Next, all valid relations are marginalized and normalized to obtain $\boldsymbol{v}^{(s,o)}$:

$$\boldsymbol{v}^{(s,o)} = \frac{\sum_{r \in \mathcal{R}_{s,o}} \boldsymbol{v}^{(s,r,o)}}{\sum_{r \in \mathcal{R}_{s,o}} \mathbf{1}^\top \boldsymbol{v}^{(s,r,o)}}, \tag{4.2}$$

where $\mathbf{1} = (1, 1, \ldots, 1) \in \mathbb{R}^{|\mathcal{X}_D|}$ is an all-ones vector. The rationale behind Equation 4.2 is that modules before the relation classifier are unaware of specific relation types. This study thus guides attention modules within the encoder to produce relation-agnostic token dependencies. Additionally, Xie et al. [105] has mentioned that in most cases, the evidence for different relation types of an entity pair turns out to be the same set of sentences.

Loss Function. The objective is to guide $\boldsymbol{p}^{(s,o)}$ with human evidence $\boldsymbol{v}^{(s,o)}$ to generate an evidence-focused localized context embedding $\boldsymbol{c}^{(s,o)} \in \mathcal{R}^d$. To achieve this, the model is trained with Kullback-Leibler (KL) Divergence loss, minimizing the statistical distance between $\boldsymbol{p}^{(s,o)}$ and $\boldsymbol{v}^{(s,o)}$:

$$\mathcal{L}_{\text{ER}}^{\text{gold}} = -D_{\text{KL}}(\boldsymbol{v}^{(s,o)} || \boldsymbol{p}^{(s,o)}). \tag{4.3}$$

During training, the balance of the effect between ER loss and RE loss is weighted using a hyper-parameter λ:

$$\mathcal{L}^{\text{gold}} = \mathcal{L}_{\text{RE}} + \lambda \mathcal{L}_{\text{ER}}^{\text{gold}}. \tag{4.4}$$

4.1.2 Weakly-Supervised Training with DREEAM

The teacher model trained on human-annotated data supports weakly-supervised DocRE and ER training on massive, unlabeled data. The data, obtained from relation distant-supervision [71], contains noisy RE labels but no ER information. A student model is trained on the data. Supervision of the student model, similar to that of the teacher model, consists of two parts: an RE loss and an ER loss.

In general, predictions from the teacher model are adopted as the supervisory signal for ER training. First, the teacher model infers on the machine-annotated data with only DocRE annotations but not evidence annotations, thereby yielding an evidence distribution over tokens $\hat{\boldsymbol{q}}^{(s,o)}$ for each entity pair (e_s, e_o). The distributions are regarded as silver evidence annotations. Next, the student model is trained to reproduce $\hat{\boldsymbol{q}}^{(s,o)}$ for each entity pair (e_s, e_o).

Loss Function. The objectives of weakly-supervised training are identical to those of supervised training. The student model is trained using a KL-divergence loss similar to Equation 4.3 to learn ER:

$$\mathcal{L}_{\text{ER}}^{\text{silver}} = -D_{\text{KL}}(\hat{\boldsymbol{q}}^{(s,o)} || \boldsymbol{q}^{(s,o)}), \tag{4.5}$$

where $\boldsymbol{q}^{(s,o)}$ is the student model's evidence distribution over tokens regarding entity pair (e_s, e_o), computed from Equation 3.3.

There are two notable differences between $\mathcal{L}_{\text{ER}}^{\text{silver}}$ and $\mathcal{L}_{\text{ER}}^{\text{gold}}$.

Firstly, the supervisory signal of $\mathcal{L}_{\text{ER}}^{\text{gold}}$ is sentence-level, while that of $\mathcal{L}_{\text{ER}}^{\text{silver}}$ is token-level. The gap results from the availability of token-level evidence distributions. On human-annotated data, it is untrivial to obtain token-level evidence distributions from sentence-level annotations automatically. On machine-annotated

data, however, the evidence distribution over tokens can be easily obtained from predictions of the teacher model. This study thus adopts token-level evidence distributions in $\mathcal{L}_{\text{ER}}^{\text{silver}}$ to provide supervision from a micro perspective for weakly-supervised ER training.

Secondly, $\mathcal{L}_{\text{ER}}^{\text{gold}}$ is computed only on entity pairs with valid relation(s), while $\mathcal{L}_{\text{ER}}^{\text{silver}}$ is computed over all entity pairs within the document. The design choice is based on the low reliability of relation labels on machine-annotated data. As these relation labels are collected automatically, some annotated relations may be underivable given the contents of the document. Therefore, it is hard to tell which relations are valid and which are not from the automatic annotations. For this reason, the loss is computed from all entity pairs to prevent missing important instances.

The overall loss is balanced by a hyper-parameter λ' similarly to that in Equation 4.4:

$$\mathcal{L}^{\text{silver}} = \mathcal{L}_{\text{RE}} + \lambda' \mathcal{L}_{\text{ER}}^{\text{silver}}. \tag{4.6}$$

Empirically, the model performs best when $\lambda' = \lambda$.

After training on the machine-annotated data, the student model is further finetuned using the human-annotated data to refine its knowledge about DocRE and ER with reliable supervisory signals. As a result, the student model is expected to outperform the teacher model because the former is trained on both human-annotated and machine-annotated data.

4.1.3 Inference

This study follows ATLOP to apply adaptive thresholding to obtain RE predictions. Specifically, relation classes with scores higher than the threshold class are selected as predictions. For ER, static thresholding is applied to choose sentences with importance higher than a pre-defined threshold as evidence.

The **inference-stage fusion** strategy proposed by Xie et al. [105] is adopted during inference, which has been detailed in Section 3.2.2.

4.2 Experiments

Experiments are conducted to evaluate the effectiveness of DREEAM.

	DocRED	Re-DocRED
# Documents	3,053 + 998	3,053 + 1,000
# Relation Types	96	96
# Relation Triples	50,455	120,664
# Relation Triples with no Evidence Annotations	1,908 (3.8%)	58,384 (48.4%)

Table 4.1: Comparison of statistics between DocRED and Re-DocRED, with the blind test set of DocRED excluded.

4.2.1 Settings

Dataset. Experiments are conducted mainly on DocRED [112][1]. As shown in Table 3.1, DocRED contains a small portion of human-annotated data and a large portion of machine-annotated data made by aligning Wikipedia articles with the Wikidata knowledge base [96]. On the other hand, the false-negative issue in the human-annotated data has been pointed out in recent studies [44, 89]. The dataset introduced to alleviate the issue is Re-DocRED, whose statistics are shown in Table 4.1. Notably, Re-DocRED supplements DocRED with relation instances but not evidence instances, making it unsuitable for training and evaluating evidence retrieval. DREEAM is also evaluated on Re-DocRED.

Computation Resources. Following previous studies, DREEAM is evaluated when using $\text{BERT}_{\text{base}}$ [28] and $\text{RoBERTa}_{\text{large}}$ [64] as the **P**retrained **L**anguage **M**odel (PLM) encoder, separately. When training and evaluating DREEAM on top of $\text{BERT}_{\text{base}}$, a single Tesla V100 16GB GPU is utilized. For $\text{RoBERTa}_{\text{large}}$, a single NVIDIA A100 40GB GPU is utilized.

Implementation. The implementation utilizes the Hugging Face's Transformers [100] library, based on the code of ATLOP [124]. However, as mentioned in Section 3.2.1, DREEAM differs from ATLOP in the computation of the token embeddings $\boldsymbol{H}$ and the cross-token dependencies $\boldsymbol{A}$. Specifically, ATLOP utilizes the last layer of the PLM encoder to compute these matrices, while DREEAM utilizes an average of the last three layers.

[1] `https://github.com/thunlp/DocRED`

Training. As mentioned in Section 4.1, a hyper-parameter λ is introduced to balance the influence between DocRE and ER training. The hyper-parameter is tuned together with the learning rate *lr* using Grid Search, where the value yielding the best performance on the development set is chosen. After searching from $\lambda \in \{0.05, 0.1, 0.2, 0.3\}$ and *lr* $\in$ {5e-5, 1e-4, 5e-4}, λ is set to 0.1 for $\text{BERT}_{\text{base}}$ and 0.05 for $\text{RoBERTa}_{\text{large}}$. After applying the same hyper-parameter search on the student model, the best performance is achieved when $\lambda = \lambda'$. When fine-tuning the student model on the human-annotated data, this study fixes λ and only searches the learning rate from {1e-6, 3e-6, 5e-6}. The optimizer is AdamW [66], with a linear warmup for the learning rate applied at the first 6% steps. Details about hyper-parameters and training time are provided in Section 4.2.4.

Evaluation. During inference, sentences x_i with $p_i > 0.2$ computed from Equation 4.1 are retrieved as evidence, where the hard threshold 0.2 is tuned based on the performance on the development set. For the evaluation, the official evaluation metrics of DocRED are adopted, namely *Ign F1* and *F1* for RE and *Evi F1* for ER [112]. *Ign F1* is measured by removing relations present in the annotated training set from the development and test sets. The reported scores are averages (and standard errors) over 5 models initialized with different random seeds.

4.2.2 Results: DocRED

Table 4.2 lists the performance of the proposed and existing methods. The best-performing model on the development set is selected to make predictions on the test set and submit the predictions to CodaLab for evaluation[2].

Performance of the Teacher Model. The upper half of Table 4.2 shows that the teacher model trained on human-annotated data exhibits comparable performance to EIDER [105] on both RE and ER. A performance gap between DREEAM and SAIS is also observed, which can be attributed to the difference in supervisory signals. While DREEAM incorporates RE with only relation-agnostic ER, SAIS is trained under three more tasks: coreference resolution, entity typing,

[2] `https://codalab.lisn.upsaclay.fr/competitions/365`. Submissions under username `kgmr15` are from this study.

Method	PLM	Dev			Test		
		Ign F1	F1	Evi F1	Ign F1	F1	Evi F1
(a) without machine-annotated data							
SSAN [106]	$\text{BERT}_{\text{base}}$	57.03	59.19	-	55.84	58.16	-
ATLOP [124]	$\text{BERT}_{\text{base}}$	59.22	61.09	-	59.31	61.30	-
E2GRE [41]	$\text{BERT}_{\text{base}}$	55.22	58.72	47.12	-	-	-
DocuNet [119]	$\text{BERT}_{\text{base}}$	59.86	61.83	-	59.93	61.86	-
EIDER [105]	$\text{BERT}_{\text{base}}$	60.51	62.48	50.71	60.42	62.47	51.27
SAIS [104]	$\text{BERT}_{\text{base}}$	59.98	62.96	53.70	60.96	62.77	52.88
DREEAM (teacher)	$\text{BERT}_{\text{base}}$	$59.60_{\pm 0.15}$	$61.42_{\pm 0.15}$	$52.08_{\pm 0.10}$	59.12	61.13	51.71
+ Inference-stage Fusion		$60.51_{\pm 0.06}$	$62.55_{\pm 0.06}$		60.03	62.49	
SSAN [106]	$\text{RoBERTa}_{\text{large}}$	60.25	62.08	-	59.47	61.42	-
ATLOP [124]	$\text{RoBERTa}_{\text{large}}$	61.32	63.18	-	61.39	63.40	-
DocuNet [119]	$\text{RoBERTa}_{\text{large}}$	62.23	64.12	-	62.39	64.55	-
EIDER [105]	$\text{RoBERTa}_{\text{large}}$	62.34	64.27	52.54	62.85	64.79	53.01
SAIS [104]	$\text{RoBERTa}_{\text{large}}$	62.23	65.17	55.84	63.44	65.11	55.67
DREEAM (teacher)	$\text{RoBERTa}_{\text{large}}$	$61.71_{\pm 0.09}$	$63.49_{\pm 0.10}$	$54.15_{\pm 0.11}$	61.62	63.55	54.01
+ Inference-stage Fusion		$62.29_{\pm 0.23}$	$64.20_{\pm 0.23}$		62.12	64.27	
(b) with machine-annotated data							
KD-DocRE [88]	$\text{BERT}_{\text{base}}$	62.62	64.81	-	62.56	64.76	-
DREEAM (student)	$\text{BERT}_{\text{base}}$	$63.47_{\pm 0.02}$	$65.30_{\pm 0.03}$	$\mathbf{55.68}_{\pm 0.04}$	63.31	65.30	**55.43**
+ Inference-Stage Fusion		$\mathbf{63.92}_{\pm 0.02}$	$\mathbf{65.83}_{\pm 0.04}$		**63.73**	**65.87**	
SSAN [106]	$\text{RoBERTa}_{\text{large}}$	63.76	65.69	-	63.78	65.92	-
KD-DocRE [88]	$\text{RoBERTa}_{\text{large}}$	65.27	67.12	-	65.24	67.28	-
DREEAM (student)	$\text{RoBERTa}_{\text{large}}$	$65.24_{\pm 0.07}$	$67.09_{\pm 0.07}$	$\mathbf{57.55}_{\pm 0.07}$	65.20	67.22	**57.34**
+ Inference-Stage Fusion		$\mathbf{65.52}_{\pm 0.07}$	$\mathbf{67.41}_{\pm 0.04}$		**65.47**	**67.53**	

Table 4.2: Evaluation results on development and test sets of DocRED, with best scores **bolded**. The scores of existing methods are borrowed from corresponding papers. The methods are grouped first by whether they utilize the machine-annotated data or not, followed by the PLM encoder.

and relation-specific ER [104]. These extra supervisory signals possibly contribute to the high performance of SAIS. Apart from the performance, the proposed method has a critical advantage over previous ER-incorporated DocRE systems in memory efficiency. A detailed discussion is provided in Section 4.3.2.

Performance of the Student Model. Table 4.2 shows that the student model outperforms existing models on RE by utilizing machine-annotated data. In particular, when adopting $\text{BERT}_{\text{base}}$ as the PLM encoder, DREEAM performs better than KD-DocRE [88], the previous state-of-the-art system, by 0.6/1.0 points on Ign F1/F1 for the development set. On the test set, the improvement reaches 1.1 F1 points on both Ign F1 and F1. Notably, DREEAM utilizing $\text{BERT}_{\text{base}}$ even performs comparably with SSAN utilizing $\text{RoBERTa}_{\text{large}}$ under the weakly-supervised setting [106]. When adopting $\text{RoBERTa}_{\text{large}}$ as the PLM encoder, the advantage of DREEAM remains on both development and test sets. These results support our hypothesis that weakly-supervised training with ER improves RE, which has not been demonstrated by any previous study.

Additionally, the student model leads the existing models by a large margin on ER. As the first approach enabling weakly-supervised ER training, DREEAM utilizes considerable amounts of data without evidence annotation via weakly-supervised training. The experimental results reveal that DREEAM improves over the state-of-the-art supervised approaches by approximately 2.0 points on Evi F1. Therefore, it is fair to conclude that the proposed approach to ER weakly-supervised training succeeds in acquiring evidence knowledge from the relation-machine-annotated data with no evidence annotation.

4.2.3 Results: Re-DocRED

As shown in Table 4.1, compared with DocRED, the training set of Re-DocRED contains many more relation triples without evidence sentences. DREEAM trained on Re-DocRED could thus be inaccurate on ER, being biased by the considerable amount of missing evidence. Therefore, evaluating ER using Re-DocRED could be sub-optimal, and this section only reports the DocRE results that are evaluated on Re-DocRED. For the same reason, during weakly-supervised training of the student model, token evidence distributions predicted by a teacher model trained on DocRED are adopted as the supervisory signal. The student model is

Method	Ign F1	F1
(a) without machine-annotated data		
ATLOP [124]	76.82	77.56
DocuNet [119]	77.26	77.87
KD-DocRE [88]	77.60	78.28
PEMSCL [37]	79.02	79.01
DREEAM	$77.34_{\pm 0.19}$	$77.94_{\pm 0.15}$
+ Inference-Stage Fusion	$79.66_{\pm 0.39}$	$80.73_{\pm 0.38}$
(b) with machine-annotated data		
ATLOP* [124]	78.52	79.46
DocuNet* [119]	78.52	79.46
KD-DocRE* [88]	80.32	81.04
DREEAM*	$78.67_{\pm 0.17}$	$79.35_{\pm 0.18}$
+Inference-Stage Fusion	$\mathbf{80.39}_{\pm 0.03}$	$\mathbf{81.44}_{\pm 0.04}$

Table 4.3: Evaluation results on the test set of Re-DocRED, with best scores **bolded**. PLM encoder is aligned to RoBERTa-large. The scores of existing methods are borrowed from Tan et al. [89]. Models with an asterisk (*) are trained using both manually and automatically annotated data.

further finetuned on Re-DocRED to obtain more reliable knowledge about RE.

Similar to Section 4.2.2, experiments are conducted under two different settings: (a) a fully-supervised setting without machine-annotated data, and (b) a weakly-supervised setting utilizing machine-annotated data. Table 4.3 compares the performance of DREEAM against existing methods. DREEAM outperforms existing methods in both fully and weakly supervised settings. As Re-DocRED is regarded as a more robust benchmark, the superiority of DREEAM on Re-DocRED further indicates its soundness.

4.2.4 Hyper-Parameters and Runtime

Important hyper-parameters are shown in Table 4.4, mainly borrowed from existing studies. Specifically, hyper-parameters are borrowed from Zhou et al. [124] to train the teacher model and from Tan et al. [88] to train and finetune the student model. The only exception is the number of epochs for training the student

	Train (teacher)		Train (student)		Finetune (student)	
Hparams.	**Bb**	**Rl**	**Bb**	**Rl**	**Bb**	**Rl**
# Epoch	30	30	2	5	10	10
lr for encoder	5e-5	3e-5	3e-5	1e-5	1e-6	1e-6
lr for classifier	1e-4	1e-4	1e-4	5e-5	3e-6	3e-6

Table 4.4: Hyper-parameters (**Hparams.**) in training. **Bb** and **Rl** represents $\text{BERT}_{\text{base}}$ and $\text{RoBERTa}_{\text{large}}$, respectively.

Phase	**$\text{BERT}_{\text{base}}$**	**$\text{RoBERTa}_{\text{large}}$**
Train (teacher)	1h18min	1h18min
Train (student)	2h55min	6h12min
Finetune (student)	26min	29min

Table 4.5: Runtime for each training stage.

model, which is determined by a grid search from $\{2, 5, 8, 10\}$.

The average running time spent for the proposed method at each training stage is shown in Table 4.5. Notably, a single Tesla V100 16GB GPU is employed when utilizing $\text{BERT}_{\text{base}}$, and a single NVIDIA A100 40GB GPU is employed when utilizing $\text{RoBERTa}_{\text{large}}$.

4.3 Analysis

4.3.1 Ablation Studies

This subsection investigates the effect of evidence-guided attention by ablation studies. All subsequent experiments adopt $\text{BERT}_{\text{base}}$ as the PLM encoder. All scores are obtained without the inference-stage fusion strategy.

Teacher Model. The first goal is to examine how guiding attention with evidence helps RE training on human-annotated data. To this end, a variant of the teacher model is trained on DocRE only, whose performance is evaluated on the DocRED development set. In general, disabling ER training reduces the model

Setting	Ign F1	F1	Evi F1
(a) Teacher Model, training documents: 3,053			
DREEAM	**59.60**$_{\pm 0.15}$	**61.42**$_{\pm 0.15}$	**52.08**$_{\pm 0.10}$
w/o ER training	59.21$_{\pm 0.19}$	61.01$_{\pm 0.20}$	42.79$_{\pm 1.65}$
(b) Student Model, training documents: 104,926			
DREEAM	**63.47**$_{\pm 0.02}$	65.30$_{\pm 0.03}$	**55.68**$_{\pm 0.04}$
w/o ER self-training	61.96$_{\pm 0.39}$	63.77$_{\pm 0.44}$	53.72$_{\pm 0.43}$
w/o ER fine-tuning	63.34$_{\pm 0.02}$	**65.50**$_{\pm 0.02}$	55.27$_{\pm 0.05}$
w/o both	62.13$_{\pm 0.07}$	63.82$_{\pm 0.08}$	47.13$_{\pm 0.12}$

Table 4.6: Ablation studies evaluated on the DocRED development set.

to a baseline similar to ATLOP [124][3].

As presented in the **Ign F1** and **F1** columns of Table 4.6a, the RE performance of DREEAM decreases without ER training. This observation supports the hypothesis that guiding attention with evidence improves RE. To investigate the effect of evidence-guided training, a case study is conducted in Section 4.3.8, where the token importance $\boldsymbol{q}^{(s,o)}$ for several instances are visualized. The visualization demonstrates that DREEAM successfully guides attention to focus more on relevant contexts.

Notably, even with the ER training completely disabled, the performance on ER still scores at an average of 42.79. The evaluation is conducted by retrieving sentences with importance higher than the pre-defined threshold, i.e., 0.2, in the ER-disabled model. This observation indicates the intrinsic task dependency between DocRE and ER.

Student Model. The second goal is to investigate the influence of training on the machine-annotated data (Equation 4.5) and finetuning on the human-annotated data (Equation 4.3). To this end, ER supervisory signals are gradually removed from the student model during the training on machine-annotated and human-annotated data. The baseline excludes ER supervision from both stages, pre-trained on machine-annotated data with only the RE loss and then finetuned on human-annotated data for only RE.

[3]The difference between ATLOP and this baseline is that the latter utilizes the last three layers of PLM to obtain embeddings, whereas ATLOP adopts only the final layer.

As shown in Table 4.6b, disabling ER training on machine-annotated data results in a huge performance drop on both DocRE and ER. This suggests that the knowledge related to ER has been effectively transferred from the teacher model to the student model via the silver evidence. On the other hand, disabling ER finetuning on human-annotated data exhibits a limited impact on performance. This is likely due to the difference in data scale. As shown in Table 3.1, the machine-annotated data is vastly larger in scale compared to the manually labeled data. Therefore, while there could be noise in the automatically assigned evidence, learning from silver evidence annotations includes a significant amount of useful information for solving the task, contributing greatly to the final performance.

Furthermore, the model trained with ER weakly-supervised training disabled (the second row in Table 4.6b) underperforms that trained with no ER loss at all (the last row in Table 4.6 (b)). The former setting is also featured with high standard deviations, with the run-scoring highest in relation to F1 outperforming that of the latter (61.96+0.39=62.35 v.s. 62.13+0.07=62.20). A possible explanation is that training the model without evidence information on the machine-annotated data has biased the attention module in a direction different from the evidence, which makes ER finetuning on gold evidence less stable and difficult to fit both sets of data. Therefore, training ER with silver evidence on machine-annotated data contributes to obtaining models with consistently high performance. These findings reaffirm the importance of weakly-supervised ER training.

4.3.2 Parameters, Memory Efficiency, and Inference Time

This subsection discusses the memory inefficiency issue in previous ER approaches and shows how DREEAM solves it.

Memory Efficiency. Previous approaches regard ER as a separate task from RE that requires extra neural network layers to solve [41, 104, 105]. To perform ER, they all introduce a bilinear evidence classifier that receives an entity-pair-specific embedding and a sentence embedding as inputs. For example, EIDER computes an evidence score for sentence x_i concerning entity pair (e_s, e_o) as in Equation 3.11. EIDER and other existing systems thus need to compute over all combinations of (sentence, entity pair).

Method	Computational Complexity	Trainable Params. (M)	Memory (GiB)
(a) without ER Module			
ATLOP [124]	$O(m^2d^2r/k)$	115.4	10.8
SSAN [106]	$O(m^2d^2r/k)$	**113.5**	**6.9**
KD-DocRE [88]	$O(m^2d^2r/k)$	200.1	15.2
(b) with ER Module			
EIDER [105]	$O(m^2d^2r/k)+O(nm^2d^2/k)$	120.2	43.1
SAIS [104]	$O(m^2d^2r/k)+O(nm^2d^2r/k)$	118.0	46.2
DREEAM (proposed)	$O(m^2d^2r/k)+O(nm^2l)$	**115.4**	**11.8**

Table 4.7: Computational complexity, trainable parameters, and memory consumption of DREEAM and existing methods. m is the number of entities, d is the dimension of token embeddings, k is the number of groups in the group bilinear classifier, r is the total size of the relation label set (97 for (Re-)DocRED), n is the number of sentences, and l is the length of the document.

Specifically, consider a document D with n sentences $\mathcal{X}_D = \{x_1, x_2, \ldots, x_n\}$ and m entities $\mathcal{E}_D = \{e_1, e_2, \ldots, e_m\}$, yielding $m \times (m-1)$ entity pairs. The longest sentence in D has l tokens in total. To obtain evidence scores, EIDER must perform bilinear classification $n \times m \times (m-1)$ times via Equation 3.11, resulting in huge memory consumption of time complexity $O(nm^2d^2)$. In actual computation, k group bilinear classifiers are adopted for approximate computation that reduces computational complexity to $O(nm^2d^2/k)$. In contrast, DREEAM takes the summations of attention weights over tokens within each sentence as evidence scores via Equation 4.1, with the time complexity being $O(nm^2l)$. As $d^2/k \gg l$, it is clear that the EIDER is more computationally costly than DREEAM[4].

Table 4.7 summarizes the computational complexity, the number of trainable parameters, and the memory consumption during inference for existing and proposed methods. The PLM encoder is aligned to $\text{BERT}_{\text{base}}$, and evaluations are made using a single NVIDIA RTX A6000 48GB GPU. Values are measured when training the models using the corresponding official repositories with a batch

[4]For experiments using $\text{BERT}_{\text{base}}$ as the encoder on DocRED, $d = 768$, $k = 16$ and $l \leq 512$.

Method	Inference Strategy	Inference Time (ms/document)	Memory (GiB)
EIDER [105]	parallel	28.1	43.1
	sequential	61.0	**7.4**
SAIS [104]	parallel	155.0	46.2
	sequential	286.0	12.7
DREEAM	parallel	**16.0**	11.8

Table 4.8: Inference time and memory consumption of existing methods and the proposed method (DREEAM). For EIDER and SAIS, the sequential inference is custom-implemented.

size of four[5]. As shown in the Table, the memory consumption of DREEAM is only 27.4% of EIDER and 25.5% of SAIS, which is a great advantage. Notably, DREEAM also consumes less memory than KD-DocRE, underscoring its memory efficiency. To clarify, the increase in memory consumption from ATLOP to DREEAM is because DREEAM utilizes the last three layers of PLM for embedding computation, while ATLOP adopts only the final layer. These observations demonstrate that, compared with existing methods, DREEAM exhibits comparable or better performance with less computation costs.

Saving Memory by Sequential Inference. On the other hand, it is possible to reduce the memory cost of existing studies by modifying the implementation and retrieving evidence of each entity pair sequentially rather than in parallel. However, for a document containing m entities, sequential inference of ER requires at least $n \times m \times (m-1)$ runs, significantly increasing the inference time. Table 4.8 shows the memory usage and inference time when performing sequential inference of ER with existing methods. While the inference time per document for DREEAM is 16.0 milliseconds, sequential inference of EIDER and SAIS takes 61.0 and 286.0 milliseconds, respectively. Here, for EIDER, the inference speed is measured on the whole document with no inference stage fusion; for SAIS, the inference time is longer as the method retrieves evidence independently for each relation label.

[5]The value of EIDER is different from the original paper because we enable ER evaluations during training.

While sequential inference reduces the memory consumption of existing methods, it significantly reduces the inference speed. Although utilizing mini-batching can accelerate the inference speed of existing methods, it is unlikely to be faster than DREEAM because mini-batch inference is slower than parallel inference. The proposed method thus strikes a better balance between time and space tradeoffs than existing methods.

Additionally, the memory consumptions of EIDER and SAIS are reduced greatly by sequential inference of ER. This suggests that the primary cause for the high memory consumption of these models is the evidence classifier, which further highlights the advantage of the proposed method in eliminating the need for an evidence classifier.

4.3.3 Evidence Retrieval and Inference Stage Fusion

Motivation. In Table 4.2 and 4.3, the performance of the proposed method is boosted after applying **I**nference **S**tage **F**usion (ISF). As introduced in Section 3.2.2, ISF is an ensemble technique that utilizes the predicted results of ER. Therefore, it is fair to hypothesize that the usefulness of ISF is positively related to the accuracy of ER. To be more specific, if the accuracy of ER is high, only the sentences that have a decisive impact on the relation extraction will be retrieved as evidence, resulting in partial documents that integrate useful information. Combining the relation predictions from these accurate partial documents with those from the original document can improve the performance of DocRE.

To verify the hypothesis, it is necessary to explore: does the performance improvement come from the information filtering effect of accurate ER? or does it simply come from the data augmentation effect from combining multiple prediction results?

Approach. This study modifies the contents of partial documents to find an answer. Specifically, for document D, three partial document sets are constructed as below:

- **Predicted Evidence** Δ: Partial documents constructed by collecting sentences in $\mathcal{V}_{s,\hat{r},o}$ from all instances $(e_s, \hat{r}, e_o, \mathcal{V}_{s,\hat{r},o})$ predicted by DREEAM. This is the real setting used in Section 4.2.

Inference Target	Ign F1	F1
Original Document D	$59.62_{\pm 0.14}$	$61.43_{\pm 0.16}$
Original Document D + Predicted Evidence Δ	$60.45_{\pm 0.13}$	$62.47_{\pm 0.16}$
Original Document D + Gold Evidence Δ^{gold}	$\mathbf{60.77}_{\pm 0.11}$	$\mathbf{62.72}_{\pm 0.21}$
Original Document D + Random Evidence Δ^{rand}	$59.70_{\pm 0.28}$	$61.81_{\pm 0.25}$

Table 4.9: Relation Extraction performance on DocRED development set of each combination of partial documents. The PLM Encoder is $\text{BERT}_{\text{base}}$.

- **Gold Evidence** Δ^{gold}: Partial documents constructed by collecting sentences in $\mathcal{V}_{s,r,o}$ from all instances $(e_s, r, e_o, \mathcal{V}_{s,r,o})$ in human annotations. This is the oracle setting where a perfect evidence retriever yields 100% correct evidence predictions. It is unlikely to obtain Δ^{gold} in practical applications.
- **Random Evidence** Δ^{rand}: Partial documents constructed by collecting random sentences. The total number of partial documents $|\Delta^{\text{rand}}|$ is aligned with that of the predicted evidence set $|\Delta|$. For each partial document $\hat{D}_i^{\text{rand}} \in \Delta^{\text{rand}}$, make sure there $\hat{D}_i \in \Delta$ such that they contain the same number of sentences $|\hat{D}_i^{\text{rand}}| = |\hat{D}_i|$. This is the setting where a model retrieves evidence randomly.

As shown in Table 4.9, ISF with random evidence Δ^{rand} yields the lowest performance. Here, as $|\Delta^{\text{rand}}| = |\Delta|$ and $\forall i, |\hat{D}_i^{\text{rand}}| = |\hat{D}_i|$, the difference in data scale between the partial document sets has been controlled, suggesting that the performance gap is due to the contents of the partial documents. This indicates that the contribution of ISF to improving DocRE is not merely due to data augmentation with partial documents.

On the other hand, ISF with gold evidence Δ^{gold} yields the highest performance. This suggests that ER with high accuracy is crucial for leveraging the effectiveness of ISF, which supports the hypothesis raised in this section. DREEAM, as the state-of-the-art model in ER, can harness the benefits of ISF, where its superiority can again be witnessed.

4.3.4 Validity of Silver Evidence

This section performs a sanity check about using token importance as silver evidence. To this end, a variant of the student model is trained under a different

Supervisory Signal	Ign F1	F1	Evi F1
Token Importance	**63.94**	**65.90**	**55.87**
Token Importance + Random Noise	63.53	65.56	55.43
Random Noise	61.98	64.10	53.34
Disabled	62.10	64.22	53.22

Table 4.10: Performance of the student model on the development set of DocRED when supervised with different silver evidence distributions.

supervisory signal during the weakly-supervised training phase. For each entity pair (e_s, e_o), the supervision of the evidence distribution over tokens $\hat{\boldsymbol{q}}^{(s,o)}$ is perturbed with a randomized noise function. The noise is generated obeying the normal distribution $\mathcal{N}(0, 1)$, after which the student model is further fine-tuned on the human-annotated data. Experiment results are shown in Table 4.10.

As in the table, when perturbed with random noise, model performance decreases for both DocRE and ER. The decrease verifies that token importance is beneficial in supervising the training of ER. When replacing the supervisory signal totally with random noise, the final performance of the model exhibits a similar level of performance to that trained with ER training disabled[6]. For relation F1, the model trained under the supervision of random noise underperforms that trained with ER supervision disabled. This observation indicates that an inefficient supervisory signal for weakly-supervised training contributes little to the final performance. In contrast, token importance improves the final model performance, demonstrating its effectiveness in supervising ER. The validity of silver evidence using token importance has thus been confirmed from this series of experiments.

4.3.5 Evaluation of Evidence Retrieval

Table 4.2 shows that the F1 score for evidence retrieval is only slightly over 50, even with DREEAM, the state-of-the-art method. While the absolute value is low, previous studies have pointed out that the evaluation metric for evidence retrieval proposed by DocRED over-penalizes models [105]. This section evaluates the performance of ER using a more appropriate metric.

[6]The scores differ slightly from those in Table 4.6 due to the randomness.

To this end, it is necessary to explain the evaluation metric of ER adopted by DocRED first. For each predicted quadruplet $(e_s, \hat{r}, e_o, \hat{\mathcal{V}}_{s,\hat{r},o})$, the performance of ER is evaluated in conjunction with that of DocRE. Specifically,

- When relation $\hat{r}$ is a correct extraction, there is a corresponding quadruplet $(e_s, \hat{r}, e_o, \mathcal{V}_{s,\hat{r},o})$ in the human annotations, with $\mathcal{V}_{s,\hat{r},o}$ being the ground-truth of ER. Therefore, the intersection $\mathcal{V}_{s,\hat{r},o} \cap \hat{\mathcal{V}}_{s,\hat{r},o}$ represents those sentences correctly extracted as evidence.

- When relation $\hat{r}$ is an incorrect extraction, it is impossible to find a corresponding quadruplet in the human annotations. Therefore, all sentences $x \in \hat{\mathcal{V}}_{s,\hat{r},o}$ are regarded as incorrect extractions.

The F1 score of ER is then computed from all predicted quadruplets. However, as the above process indicates, ground-truth evidence is only defined for ground-truth relations. As a result, two kinds of errors will occur corresponding to the error of relation extraction:

- If a model predicts a wrong relation $\hat{r}$ between an entity pair e_s, e_o, then all sentences in the evidence $\hat{\mathcal{V}}_{s,\hat{r},o}$ corresponding to $\hat{r}$ are regarded as incorrect. This contributes to false-positive predictions, decreasing the precision of ER evaluation.

- If a model fails to detect a relation r between an entity pair e_s, e_o, the evidence $\mathcal{V}_{s,r,o}$ will not be included in the evaluation of ER. This contributes to false-negative predictions, decreasing the recall of ER evaluation.

In short, the evaluation of ER is highly in conjunction with that of DocRE: Models with higher performance on DocRE tend to score higher on ER as well. The evaluation metric of ER adopted in DocRED, therefore, is unsuitable for measuring purely the performance of ER.

To separate the impact of DocRE from the evaluation of ER, Xie et al. [105] proposed **PosEvi F1**. PosEvi F1 focuses solely on relation triples $(e_s, \hat{r}, e_o)$ that are correctly predicted by the model and measures how accurately their evidence $\mathcal{V}_{s,\hat{r},o}$ is retrieved. Relation triples wrongly predicted by the model, i.e., false-positives, and relation triples in the ground-truth annotation but failed to be extracted by the model, i.e., false-negatives, are ignored during the evaluation.

	Evi F1			PosEvi F1		
	Pre.	**Rec.**	**F1**	**Pre.**	**Rec.**	**F1**
EIDER	52.72	**46.56**	49.72	83.24	**83.19**	83.21
DREEAM	**59.63**	46.25	**52.10**	**92.03**	82.94	**87.25**

Table 4.11: Evidence Retrieval performance on DocRED development set when using different evaluation metrics. **Pre.** and **Rec.** represent Precision and Recall, respectively. The PLM Encoder is $\text{BERT}_{\text{base}}$.

	Rel Ign F1 (ISF)	Rel Ign F1	Evi F1
0(0%)	59.98	59.14	43.30
30(1%)	60.45	59.32	48.85
305(10%)	60.28	59.46	50.92
915(30%)	60.45	59.49	51.64
1,526(50%)	60.45	59.49	51.89
2,137(70%)	60.33	59.45	51.91
2,747(90%)	60.45	59.52	52.05
3,053(100%)	60.77	59.62	52.10

Table 4.12: Performance of DREEAM on DocRED development set when varying the number of documents used for ER training.

The results are shown in Table 4.11, where the performance of EIDER is based on reproducibility experiments.

Table 4.11 demonstrates that both EIDER and DREEAM score higher than 80 when evaluated with PosEvi F1. The observation showcases the capability of both methods in retrieving correct evidence for correctly-predicted relation triples. However, even when utilizing DREEAM for ER, the F1 score is lower than 90, which shows the difficulty of perfect evidence retrieval. Notably, while the performance gap between EIDER and DREEAM is 2.38 when evaluating with Evi F1, the gap goes up to 4.04 when evaluating with PosEvi F1. Therefore, DREEAM, the proposed method, further exhibits its superiority in retrieving evidence for correctly predicted relation triples.

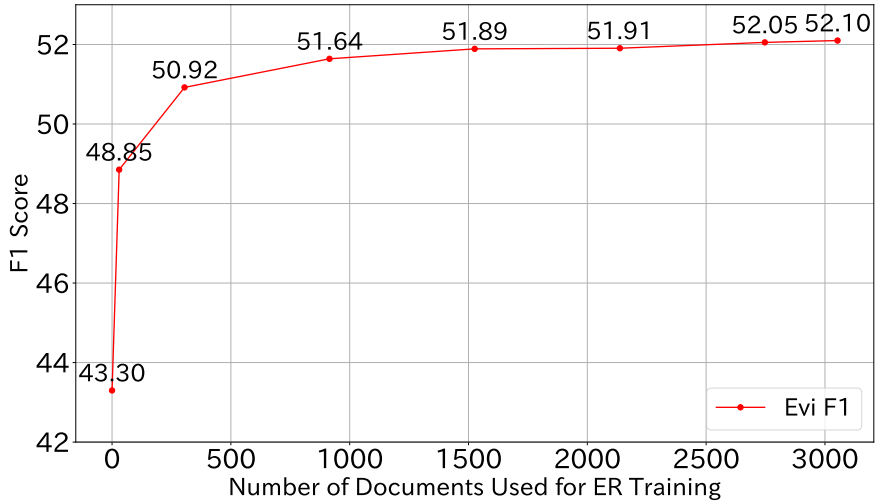


Figure 4.3: Evidence F1 of DREEAM when varying the number of documents used for ER Training.

4.3.6 Amount of Data For Training Evidence Retrieval

As shown in Section 4.3.1, training ER helps improve DocRE. Section 4.3.3 confirms that high-performance ER can harness the benefits of Inference Stage Fusion. However, it is costly to manually annotate evidence, resulting in a limited number of human annotations. This section investigates the relationship between the number of training instances used for ER and the model performance. The aim is to determine the minimum data required to retrieve evidence of DocRE fairly accurately.

Figure 4.3 and Table 4.12 report DREEAM's performance in ER when varying the number of documents used for ER training. Note that for DocRE, all available training documents are utilized, and ISF is conducted during inference. The PLM is $\text{BERT}_{\text{base}}$.

The experiment results show a significant improvement in ER performance as the number of documents used for ER training increases from 0 to 1,526 (i.e., 50% of the total training set). However, when further increasing the training data beyond 1,526 documents, the performance improvement is subtle, eventually stabilizing at almost the same level[7]. This suggests that DREEAM can retrieve

[7]Note that although the setting of using all documents for ER training aligns with that in Section 4.2.2, the reported value differs due to the randomness.

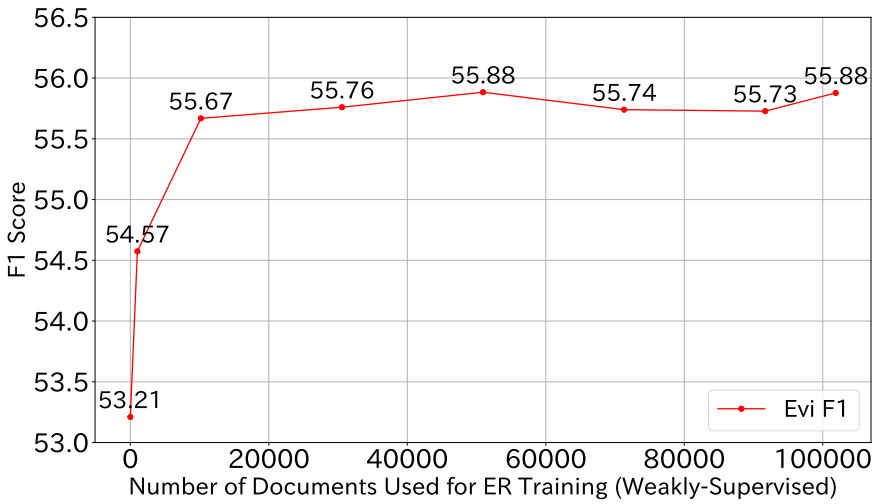


Figure 4.4: Evidence F1 of DREEAM when varying the number of documents used for ER Training.

evidence with reasonable accuracy even with reduced training data. Additionally, the greatest performance improvement can be observed when increasing the training data from 0 to 30 documents (i.e., 1% of the total training set). The observation indicates that guiding attention with evidence is effective even with a tiny training data set.

It is also interesting to investigate how the amount of data used in the weakly supervised setting influences the final performance of ER. To this end, a similar series of experiments are conducted where instances utilized for ER training are limited during weakly-supervised training. Specifically, modifications are made in Step (3) of Figure 4.2, where only a pre-determined fraction of the training instances with silver evidence contribute to ER training. The student model is further finetuned using the human-annotated data as in Step (4). Evaluation results are plotted in Figure 4.4. From the figure, a similar tendency can be observed that increasing the number of training instances benefits performance the most when the amount of resources is limited. This observation supports that the proposed weakly-supervised strategy can effectively utilize unlabeled data for ER training, even when the number of instances is limited. Additionally, the tendency in Figure 4.4 is less stable compared to that of Figure 4.3, indicating that supervision using gold annotations is more reliable than that using silver annotations.

Based on the above analysis, it can be confirmed that DREEAM can achieve a certain level of accuracy even when the amount of data for ER is limited, under both fully-supervised and weakly-supervised settings.

4.3.7 Evidence Distribution of Triples

DREEAM and other methods tackling DocRE and ER jointly assume that evidence annotations provide a meaningful set of sentences $\mathcal{V}_{s,r,o}$ that indicate the relation triple (e_s, r, e_o). A trivial heuristic for deciding the evidence is to select all sentences explicitly mentioning entities e_s and e_o. Such a heuristic results in two kinds of errors:

- **False-Positives**: Non-evidence sentences that mention the subject or object entity are included in the evidence set;
- **False-Negatives**: Evidence sentences that do not mention both the subject and object entities are excluded from the evidence set.

To provide a concrete example, when considering the triple demonstrated in Figure 1.3, a trivial evidence annotation will be {1,2,5,6}, where sentence 1 contains the mention of the object entity *Blackadder* and sentences 2,5,6 contain the mentions of the subject entity *Prince Edmund*.

While such risks exist in collecting evidence annotations, no information is provided in the original paper of DocRED describing how they are mitigated during the annotation process [112]. If all annotators follow this heuristic for annotating evidence, the annotations will deviate from the initial goal of collecting evidence annotations, which is to assist in locating information necessary for relation extraction. On the other hand, if a model merely follows this heuristic for predicting evidence, it should be considered sub-optimal since it fails to filter information effectively. This section looks into the distribution of manually annotated and predicted evidence sentences to evaluate the effort exerted by annotators and models in evidence retrieval.

Positive Evidence. This study divides positive evidence into two types:

- **Explicit Evidence** defined as evidence sentences that explicitly mention the corresponding subject or object entity;

	Explicit Evidence			Implicit Evidence		
	Total	**True**	(%)	**Total**	**True**	(%)
Annotated	18,133	18,133	(100%)	1,324	1,324	(100%)
EIDER	15,436	5,831	(37.78%)	1,586	65	(4.62%)
DREEAM	14,778	6,825	(**46.18%**)	323	36	(**11.15%**)

Table 4.13: Number of intra and extra evidence sentences from human annotations and model predictions.

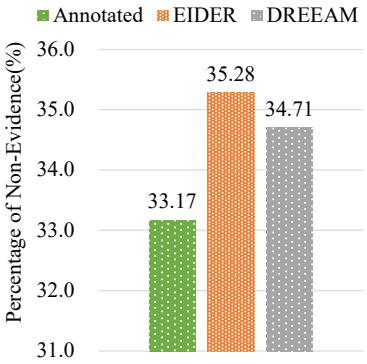


Figure 4.5: Percentage of non-evidence sentences mentioning corresponding entities on the development set of DocRED.

- **Implicit Evidence** defined as evidence sentences that do not mention the corresponding subject or object entity.

Table 4.13 shows the statistics of explicit and implicit evidence of human annotations and model predictions. EIDER and the proposed method DREEAM are selected as the representatives. Several observations can be made from the table. Firstly, human annotations yield significantly more explicit than implicit evidence. The result indicates that, in most cases, relation extraction decisions are based on explicit rather than implicit evidence. Secondly, the accuracy of model predictions on implicit evidence is significantly lower than on explicit evidence. This demonstrates the difficulty of identifying implicit evidence. Finally, compared with EIDER, DREEAM yields more accurate predictions for both explicit and implicit evidence. The finding underscores the superiority of DREEAM in evidence retrieval.

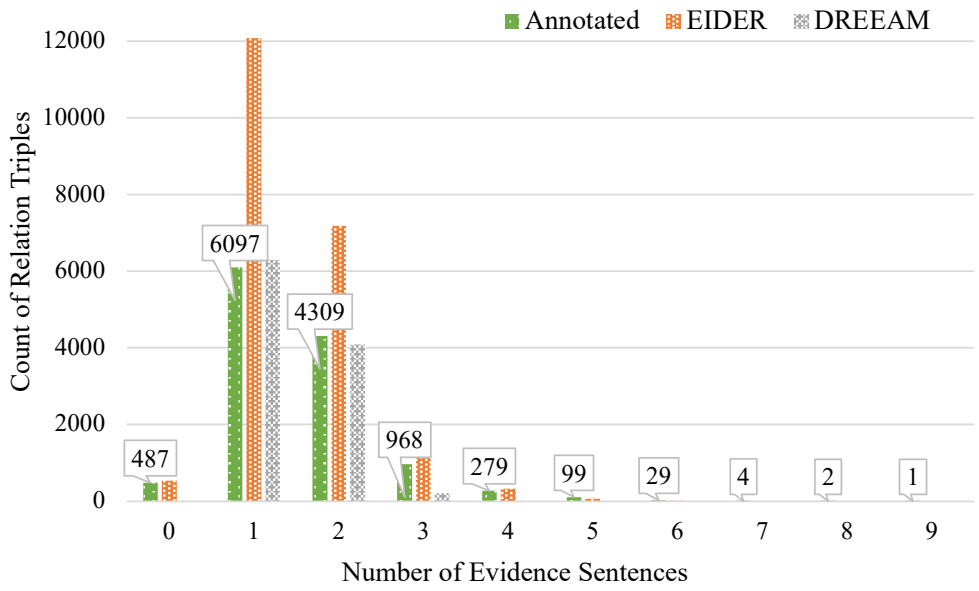


Figure 4.6: Evidence distribution of relation triples on the development set of DocRED. Only the number of relation triples in human annotations are labeled in the figure for clarity.

Negative Evidence. Next, this study examines how many non-evidence sentences containing a corresponding entity mention are removed in human annotation and model predictions, with results shown in Figure 4.5. For each triple (e_s, r, e_o), the figure reports the percentage of sentences that mention e_s or e_o but are not included in the evidence set $\mathcal{V}_{s,r,o}$, relative to all sentences mentioning e_s or e_o.

In both human annotations and model predictions, more than 30% of the sentences mention corresponding entities but are excluded from the evidence. This suggests that the false-positive problem mentioned earlier in this section may not be severe.

Finally, Figure 4.6 depicts the distribution regarding the number of evidence sentences in human annotations and model predictions. The number of evidence sentences in human annotations ranges from 0 to 9, primarily concentrated between 0 and 3. Model predictions share a similar tendency as the human annotations, while EIDER yields significantly more relation triples with only 1 evidence sentence. Notably, some relation triples have an empty set of evidence in human

annotations. This observation applies to predictions from EIDER but does not hold true for DREEAM. Examining relation triples with an empty evidence set in human annotations reveals a common characteristic: they often pertain to common sense. For example, the relation (Indiana State Teachers College, *located in*, Indiana) can be inferred from only the lexical information as they share the same word "Indiana". Therefore, annotators might have left the evidence empty, assuming that no additional information from the document was necessary.

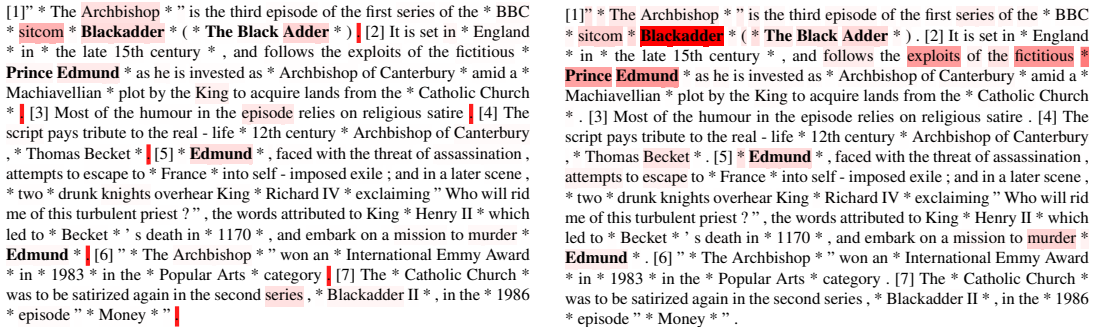
4.3.8 Visualization: Evidence-Guided Attention

This section investigates the influence of the proposal: guiding attention with evidence from case studies. As introduced in Section 4.1.1, evidence knowledge of DREEAM originates from sentence-level supervision. This study hypothesizes that sentence-level supervision, from a more macro perspective, should improve its micro counterpart of token-level focusing. To test the hypothesis, this study looks into the token-level importance distribution for localized context pooling $\boldsymbol{q}^{(s,o)}$, computed from Equation 3.3, before and after evidence-guided training. Specifically, $\boldsymbol{q}^{(s,o)} \in \mathcal{R}^{|\mathcal{T}_D|}$ with and without evidence-guided training are visualized as heatmaps using the toolkit developed by Yang and Zhang [110]. The distribution without evidence-guided training is obtained from a variant of DREEAM where the hyper-parameter controlling the influence of ER training, i.e., λ in Equation 4.4, is set to 0.

Cases when the extracted relation is correct. Figure 4.7 shows an example of heatmaps when the relation is correctly predicted by DREEAM.

Firstly, when focusing on the sentence level, tokens in the first and the second sentences are assigned with higher weights after attention guidance. Given that the ground-truth evidence is sentences 1 and 2, it is fair to conclude that the token importance $\boldsymbol{q}^{(s,o)}$ has been appropriately guided by the sentence-level teacher signal $\boldsymbol{v}^{s,o}$ as in Equation 4.3.

Secondly, when focusing on the token level, it is clear that before training the evidence-guided attention, the model tends to focus on the period of each sentence. Guiding the attention with evidence helps the model to focus more on the critical tokens providing a clue for relation classification, e.g., *fictitious* in Figure 4.7b that are related to the relation *present in work*. Notably, $\boldsymbol{v}^{s,o}$ is a



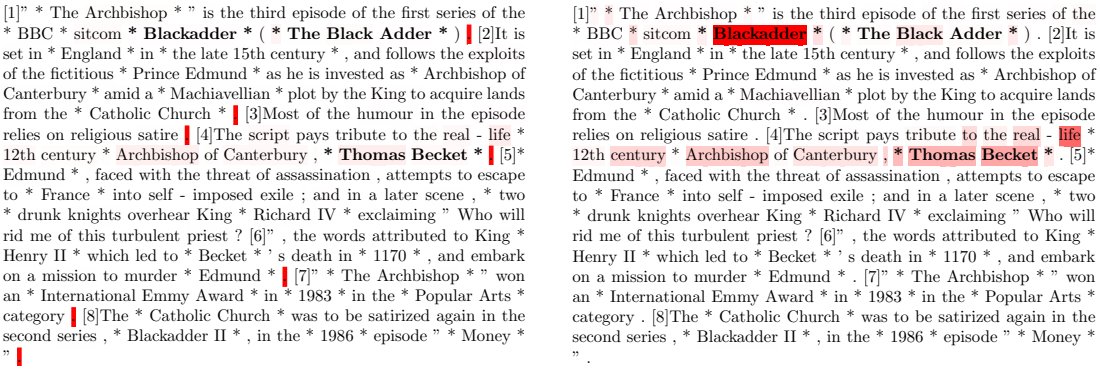
(a) Before attention guidance. (b) After attention guidance.

Figure 4.7: Heatmaps of token importance for localized context pooling before and after guiding the attention with evidence when deciding the relation for entity pair (*Prince Edmund*, *The Black Adder*). The gold relation is *present in work* with evidence sentences 1 and 2. The deeper the color, the larger the value.

sentence-level supervisory signal computed from sentence-level human annotations, which is not designed to increase the weight of particular words or tokens. Despite this, the word-level importance distribution has been successfully guided to meaningful words, suggesting that the sentence-level teacher signal indirectly guides the model to focus on words important for relation classification. On the other hand, a model without the guidance of evidence assigns high weights to the end of each sentence, i.e., the period, which differs from human decision-making criteria. The same phenomena have also been observed by [14], a study that investigates whether the model understands the document based on evidence retrieval. Therefore, guiding attention with evidence aligns the decision-making criteria of the model with humans, improving the explainability of models.

Cases when the extracted relation is incorrect. Figure 4.8 shows an example of heatmaps when the relation is wrongly predicted by DREEAM.

As mentioned in Section 4.3.5, evidence is not defined for entity pairs with no relations. However, after training with evidence-guided attention, the model assigns high weights on the word *life* and yields a prediction *author* between entity pairs (*Thomas Becket*, *The Black Adder*). While it is possible to correlate the word *life* with the relation label *author* in phrases "life-long work" or "lifetime



(a) Before attention guidance.

(b) After attention guidance.

Figure 4.8: Heatmaps of token importance for localized context pooling before and after guiding the attention with evidence when deciding the relation for entity pair (*Thomas Becket*, *The Black Adder*). The deeper the color, the larger the value. There is no relation between the entity pair in human annotations, but the model predicts relation *author* with evidence sentences 1 and 4.

work", such a context cannot be observed in this document. This suggests that merely increasing the importance of sentences or tokens serving as evidence may not explicitly account for context or semantics. As a result, the model might learn a simplistic correlation between relation labels and tokens, potentially leading to incorrect predictions. Therefore, designing an evidence retrieval method that considers contextual information remains a future challenge.

4.3.9 Error Analysis for ER

This section investigates cases in which DREEAM makes wrong predictions. As the previous section has introduced error cases for RE, this section conducts error analysis with a special focus on ER. As mentioned in Section 4.3.5, measuring the vanilla F1 scores for ER is sub-optimal as it penalizes the performance with the accuracy of RE. This section follows Section 4.3.5 to investigate the performance of ER based on PosEvi F1. Specifically, for each correctly-predicted relation triple $(e_s, \hat{r}, e_o)$, the accuracy of its evidence $\hat{\mathcal{V}}_{s,\hat{r},o}$ is measured by comparing against the manual annotation $\mathcal{V}_{s,\hat{r},o}$.

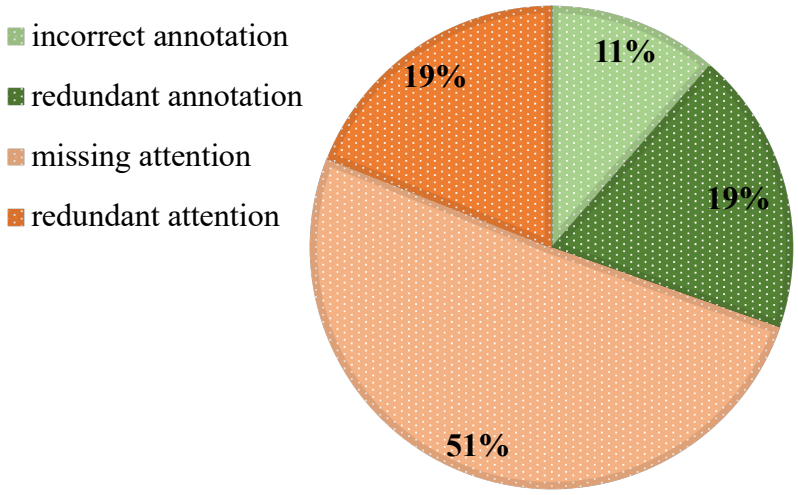


Figure 4.9: Error analysis for ER on 100 randomly-sampled error cases when predicting on the development set of DocRED.

Here, 100 error cases with a correctly-predicted relation label ($\hat{r} = r$) and a wrongly-retrieved evidence sentence set ($\hat{\mathcal{V}}_{s,\hat{r},o} = \mathcal{V}_{s,\hat{r},o}$) are randomly sampled from model predictions. The error cases are manually investigated are classified into five groups:

- **incorrect annotation**: The evidence annotation is considered to be incorrect, with necessary evidence sentence(s) missed to decide the relation label;
- **redundant annotation**: The evidence annotation is considered to be redundant, with a subset of which is enough for deciding the relation label;
- **missing attention**: The model fails to assign high attention weights to evidence sentences, resulting in missing evidence predictions.
- **redundant attention**: The model assigns high attention weights to unnecessary sentences, resulting in redundant evidence predictions;

The fraction of each group is demonstrated in Figure 4.9, with the PLM encoder being $\text{BERT}_{\text{base}}$.

(a) missing attention
[1] **Aino** is a figure in the Finnish national epic Kalevala.
[2] It relates that she was the beautiful sister of **Joukahainen** .
[3] Her brother, having lost a singing contest to the storied Vinminen, promised **Aino**'s "hands and feet" in marriage if Vinminen would save him from drowning in the swamp into which **Joukahainen** had been thrown.
[4]...
Relation: (Joukahainen, sibling, Aino) **Evidence** (gold/pred): [1,2]/[2]
(b) redundant attention
[1] **Allen** **County** is a county in the U.S. state of Ohio.
[2] As of the 2010 census, the population was 106,331.
[3] The county seat is **Lima** .
[4] ...
[7] **Allen** **County** comprises the **Lima** , OH Metropolitan Statistical Area , which is also part of the **Lima**-Van Wert-Wapakoneta, OH Combined Statistical Area.
[8]...
Relation: (Allen County, capital, Lima) **Evidence** (gold/pred): [1,3]/[1,3,7]

Table 4.14: Case studies for prediction errors of ER from DREEAM.

From the pie graph, it can be observed that the most frequent error type is missing attention, i.e., the model failing to focus on necessary texts. A typical case is shown in Table 4.14(a). Here, the model assigns high weights to "sisters" but forgets about the head and tail entities, resulting in mispredictions of evidence for the correctly predicted relation label. This observation further supports our hypothesis in Section 4.3.8 that guiding attention with evidence might result in the model learning a simplistic correlation between relation labels and tokens. The model might thus overlook some sentences that function as evidence in the reasoning chain when deciding relations. An example of redundant attention is also showcased in Table 4.14(b). In the example, while attention to the first and the third sentences is enough to classify the relation, the model still picks up the seventh sentence as evidence. The example demonstrates that the proposed method does not guarantee a minimal evidence set.

Additionally, redundant and missing annotations also comprise a non-neglectable

portion in Figure 4.9. The observation indicates that there is still room for improvement in controlling the quality of evidence annotations.

4.4 Summary

This chapter has introduced DREEAM, the method proposed in the dissertation to improve the usage of ER in DocRE, and how to utilize the model for training a better DocRE and ER model.

DREEAM is a memory-efficient model that directly incorporates the ER supervisory signal into the relation classifier. Unlike existing approaches that train an evidence classifier for ER, DREEAM directly supervises the attention to concentrate more on evidence than on others. DREEAM can be employed in a weakly-supervised setting to compensate for the shortage of human annotations. Instead of gold evidence annotated by humans, evidence predictions from a teacher model trained on human-annotated data are adopted as the supervisory signal to realize weakly-supervised ER training on unlabeled data.

Experiments on two widely-used benchmarks, namely the DocRED and Re-DocRED, show that DREEAM exhibits state-of-the-art performance on both DocRE and ER, with the help of weakly supervised training on data obtained from distant supervision of relations. Compared with existing approaches, DREEAM performs ER with introduced zero trainable parameters, thereby reducing memory usage to 27% or less.

Analyses have been conducted detailing the computation efficiency of DREEAM. The proposed method also exhibits its superiority in an improved level of explainability compared to existing studies.

5 Dataset Construction: JacRED

DocRE research has been conducted mainly in English [89, 112, 124]. However, a need exists to automatically construct Knowledge Bases (KBs) in non-English languages: Every language has its own monolingual documents containing language-specific knowledge unavailable in English data. Given that KBs are attracting increasing attention as a data source to alleviate hallucinations in LLMs [1, 6], it is urgent to design methods for efficiently collecting data for training non-English DocRE models. This study thus explores ways to construct language resources for non-English DocRE with reduced human efforts, which can further serve as the basis for training non-English DocRE models.

Specifically, this study utilizes existing resources of English DocRE to construct datasets and models for non-English DocRE. Japanese is chosen as the target language for the following two reasons:

- Despite Japanese being a widely used language for web content, there is currently a notable absence of general-purpose Japanese DocRE resources. This study thus contributes to the community by establishing the foundation for Japanese DocRE.
- Secondly, Japanese stands out as one of the most linguistically distant languages from English [19]. The dissimilarity encompasses various aspects, including script models and word order. Therefore, this research setting is highly representative, and the insights will hold value when acquiring resources for other languages.

This study first explores if DocRE resources of high quality can be obtained with *zero human effort*. To this end, a Japanese DocRE dataset is automatically constructed with cross-lingual projection (Section 3.3.3). Specifically, Re-DocRED [89], a popular English DocRE dataset of high quality, is translated into Japanese with a machine translator. An automatically constructed dataset

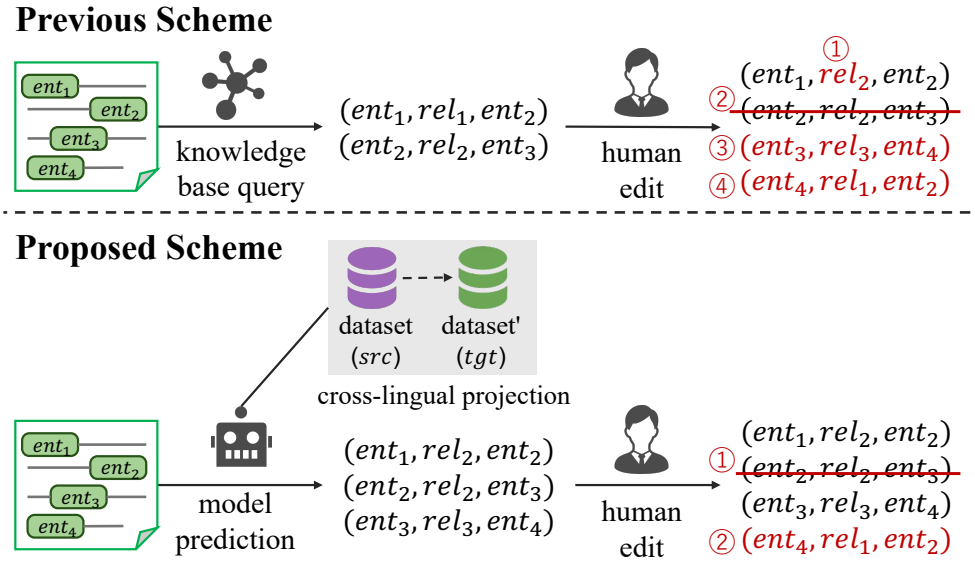


Figure 5.1: Overview of the proposed annotation scheme. *src* and *tgt* represent the source and target language, respectively. The existing scheme requires 4 human edit steps to reach the final annotation, while the proposed method only requires 2.

(hereafter referred to as Re-DocRED^{ja}) can thus be obtained without human annotators. The translation-based cross-lingual projection has been successfully applied to other information extraction (IE) tasks, including named entity recognition and sentence-level relation extraction [16, 40].

However, models trained on Re-DocRED^{ja} suffer from low recalls when extracting relation triples from "real" Japanese text, i.e., those written by native speakers. After investigating the error cases, the failures can be attributed to the discrepancies between documents in Re-DocRED^{ja} and those composed by native speakers, as well as the error propagation from the machine translator. These observations underscore the uniqueness and complexity of DocRE in comparison to other IE tasks.

Given that Re-DocRED^{ja} is unsuitable for practical application, the next step is to explore whether the dataset can assist human annotation. As in Figure 5.1, this study adopts a *semi-automatic, edit-based annotation scheme*, where annotators edit machine recommendations by removing incorrect instances and supplementing missed instances [18, 89, 112]. In contrast to previous studies where

Dataset	Lang.	# Triples	# Docs.	Avg. # Toks.	# Rels.	Evi.
DocRED [112]	*en.*	50,503	4,051	198.4	96	Y
Re-DocRED [89]	*en.*	120,664	4,053	198.4	96	N
HacRED [18]	*zh.*	56,798	7,731	122.6	26	N
HistRED [111]	*kr.*	9,965	5,816	100.6	20	Y
JacRED	*ja.*	42,241	2,000	260.1	35	Y

Table 5.1: Statistics of existing and proposed DocRE datasets. Column **Evi.** shows whether each dataset annotates evidence sentences or not. Statistics for DocRED are from the human-annotated subset.

only relation instances from an existing KB are recommended [18, 112], instances in this study are recommended by a state-of-the-art DocRE model trained on Re-DocRED^{ja}. The constructed dataset is named **JacRED** (Japanese Document-level Relation Extraction Dataset), with statistics shown in Table 5.1. Quantitative analyses are conducted on recommendations from the model trained on Re-DocRED^{ja} and those from knowledge base queries, where the former reduces the human edit steps to half of the latter.

JacRED can be adopted as a benchmark to assess DocRE models. When evaluating the performance of existing models on Japanese DocRE, models trained using the train set of JacRED perform fairly well on the test set, while the scores fall short of those achieved on Re-DocRED. The result indicates that JacRED introduces extra challenges to Re-DocRED. Notably, in-context learning of LLMs yields poor performance on JacRED, in line with the findings of Wadhwa et al. [98].

Quantifying the performance gap between models trained on Re-DocRED^{ja} and those trained on JacRED, the former falls short of the latter to around 10 F1 points. The results suggest that although translation-based cross-lingual projection effectively constructs multilingual datasets for a range of IE tasks, it does not hold true for DocRE.

Additionally, JacRED also enables the evaluation of cross-lingual DocRE. Model's cross-lingual transferability between English and Japanese can be assessed using JacRED and Re-DocERD, from which challenges can be observed due to the complexity of document semantics.

In short, the contributions of the dataset construction part are:

- The proposal of the cross-lingual projection-based scheme for constructing DocRE language resources. The proposed method reduces necessary human annotation steps by half compared with existing annotation methods.
- The inspection on automatically constructing DocRE datasets with existing translation-based cross-lingual projection approach and how it fails. The findings portray the difference between DocRE and other IE tasks.
- The publishment of JacRED, the first general-purpose DocRE dataset for Japanese. JacRED not only sets up the foundation of Japanese DocRE research but also enables cross-lingual evaluation of DocRE tasks.

In the following chapter, Section 5.1 details the process of data construction, Section 5.2 investigates the statistics of JacRED as a representative constructed using the proposed dataset-construction approach, and Section 5.3 reports experiment results using the newly-constructed dataset JacRED.

5.1 Dataset Construction Method

5.1.1 Strategy

The aim is to construct a Japanese DocRE dataset ready for use with minimal human effort. To this end, the first attempt is to build a dataset without human annotators automatically. The approach has been reported successful in other IE tasks [16, 40].

If the automatically constructed dataset is qualified to train Japanese DocRE models, there is no need to recruit human annotators. However, if the quality of the automatically constructed dataset is unsatisfactory, the next attempt is to use the dataset as an intermediary tool to assist human annotation. To this end, this study employs models trained on the automatically constructed dataset to recommend relation triple candidates, based on which human annotators make revisions to reach the final annotations.

The rest of this section details each of the attempts in Section 5.1.2 and 5.1.3, respectively.

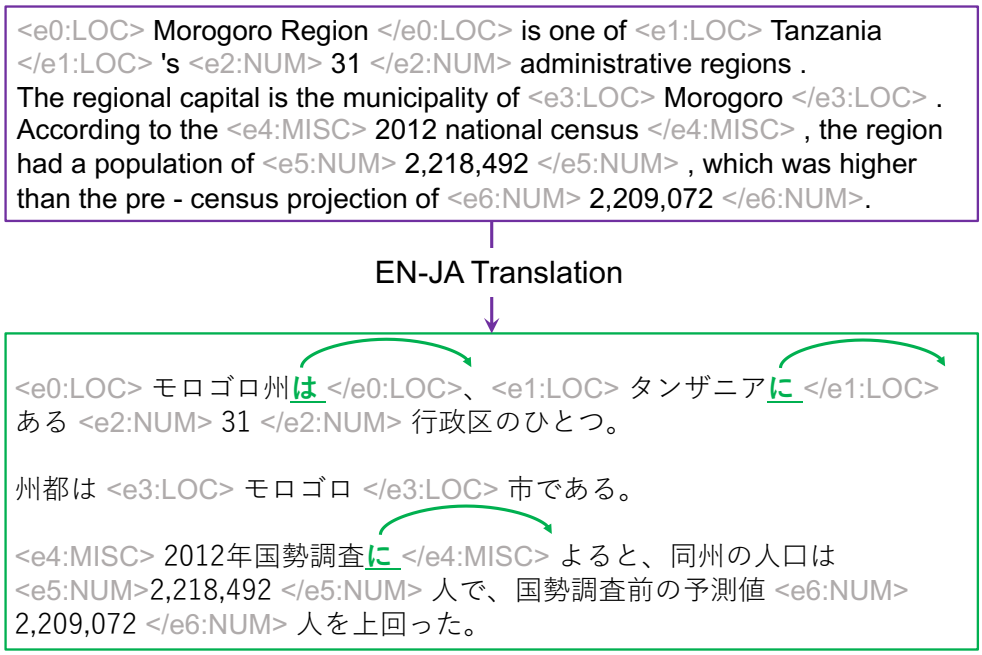


Figure 5.2: Translating Re-DocRED from English into Japanese with label projection. Translations went through post-edits to detach case markers from entity spans.

5.1.2 Automatic Construction

This study first builds a Japanese version of Re-DocRED [89]. Re-DocRED is a revised version of DocRED [112], the first and most popular DocRE dataset constructed from English Wikipedia. Details about these datasets are introduced in Section 3.3.1.

Translation and Annotation Projection. The complete train/dev/test splits of Re-DocRED are translated into Japanese with the help of machine translators. As shown in Figure 5.2, XML tags are adopted as entity span markers and inserted around each entity, following Chen et al. [16] and Hennig et al. [40]. Documents are translated from English to Japanese while preseving the tags so that entity spans are projected jointly during translation. Relations associated with entity pairs can thus be directly inherited from the English dataset. The machine translator is DeepL, which enables translation while preserving XML

tag markups[1].

Post-processing for Case Markers. The translated Japanese sentences, as in Figure 5.2, require post-editing due to the presence of case markers in entity spans. Case markers ("kaku-joshi" in Japanese) are special linguistic units attached to the end of nouns to indicate the relationship between words. A case marker only reveals the grammatical role but does not contribute to the semantics of the noun phrase it is attached to. For example, in entity span **<e0>** of the Japanese translation, a topic marker "は" following "モロゴロ州" (Morogoro Region) indicates the noun phrase to be the topic of this sentence.

Case makers from the entity span are detached with the Japanese morphological analyzer MeCab [53][2]. To this end, tokens identified as case markers at the end of each entity span are detached and re-attached at the corresponding position outside the span. The obtained dataset is denoted as Re-DocRED^{ja}.

Limitations of the Translated Dataset. When utilizing Re-DocRED^{ja} as the training data and test bed, DREEAM [70], the current state-of-the-art DocRE model proposed in Chapter 4, achieves an F1 score of 72.74 (cf. the same architecture scores 77.94 on the original Re-DocRED). However, when "real" Japanese documents from Japanese Wikipedia are fed into the model, quite some relation triples are left out in the predictions, with a typical example showcased in Figure 5.3. Two possible reasons can be raised to explain why the model trained on Re-DocRED^{ja} fails:

1. **Topic Shift of Contents**: Re-DocRED^{ja} cannot represent the real topic distribution of Japanese documents. Constructed from English Wikipedia, Re-DocRED consists of contents that English speakers are concerned about, which do not necessarily match the interests of Japanese speakers. As in Figure 5.3a, Re-DocRED^{ja} lacks documents about Japanese culture, preventing the DocRE models from being localized.

2. **Difference in Surface Form**: The surface structures, i.e., how words are organized in the sentence, of Re-DocRED^{ja} somehow follow the logic of English, which is distinct from that of Japanese. Figure 5.3b showcases a

[1] `https://api.deepl.com/v2/translate`
[2] `https://taku910.github.io/mecab/`

JA: 堀 直宥(ほり なおさだ、寛文5年11月17日(1665年12月23日) - 正徳元年6月8日(1711年7月23日))は、江戸時代前期から中期の大名で、上総八幡藩第2代藩主。

EN: Naosada Hori (December 23, 1665 - July 23, 1711) was a feudal lord of the early to mid-Edo period, the second lord of the Joso Hachiman domain.

missed triple: (Naosada Hori, head of government, Joso Hachiman domain)

(a) Example of unextracted relations due to the topic shift of contents. The highlighted "藩主" is a Japanese historical term used from 1603 to 1912 meaning "lord".

JA: ザカリアーシュ・ヨージェフ(1924年3月25日 - 1971年11月22日)は、ハンガリー出身のサッカー選手、サッカー指導者。
1954年のFIFAワールドカップでは決勝戦を除く4試合にフル出場し準優勝に貢献した。

EN: Zakarias Yogev (March 25, 1924 - November 22, 1971) was a Hungarian soccer player and soccer coach.
(He) played in all but the final four games of the 1954 FIFA World Cup, contributing to the runners-up finish.

missed triple: (Zakarias Yogev, participant in, the 1954 FIFA World Cup)

(b) Example of unextracted relations due to the gap in surface structures. The subject of the second sentence is left out in Japanese.

Figure 5.3: Cases where the model trained on Re-DocRED^{ja} failed to predict. Documents are shown as partial for better visibility. Note that English translations are provided only for reference, while predictions are actually made in Japanese texts.

typical example of how Japanese differs from English in surface structures regarding the omission of subjects. Re-DocRED^{ja} thus cannot reproduce the surface structures of "real" Japanese, resulting in failures of the trained model.

5.1.3 Semi-Automatic Construction

Drawbacks of Re-DocRED^{ja} postulate that human annotations are necessary to depict Japanese DocRE better. This study thus involves human annotators in constructing a Japanese DocRE dataset. Following the scheme provided by Yao et al. [112], the annotation process consists of two phases: the entity mention annotation phase and the relation annotation phase. Both phases follow an edit-based scheme, as shown in Figure 5.4: Annotators only need to edit machine

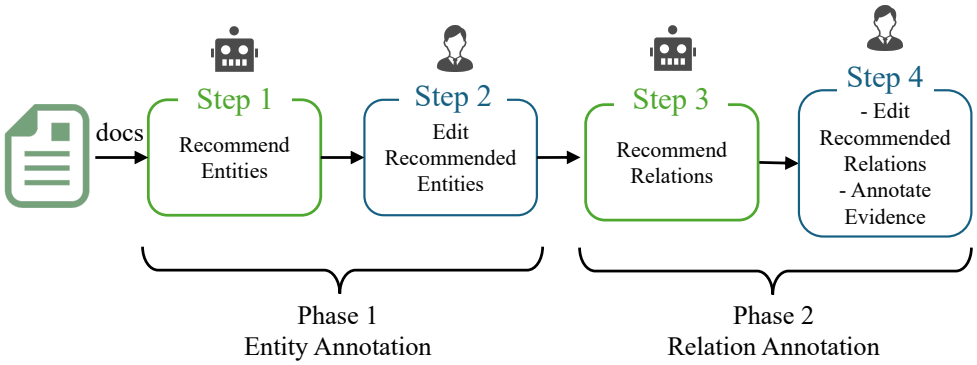


Figure 5.4: The annotation pipeline used in Yao et al. [112], which is also adopted by this study. This study makes two proposals in Step 3.

recommendations instead of enumerating all relation instances from scratch.

The quality of machine recommendation is crucial under the edit-based scheme: Poor recommendations require more edits, drastically increasing the annotators' workload and affecting the dataset's quality. The problem is recognized in DocRED as the *false-negative issue* (Section 3.3.1), where too many relation instances are left out in the recommendations to be mended by human edits [44]. This study proposes to mitigate this issue using Re-DocRED^{ja}, utilizing a model trained on Re-DocRED^{ja} to recommend relation instances. The rationale is that Re-DocRED is an improved version of DocRED aiming at alleviating the false-negative issue; hence, a model trained on Re-DocRED^{ja}, the translated version of Re-DocRED, could be expected to inherit the advantage and yield considerably more predictions.

Documents. JacRED is built on top of the Japanese edition of Wikipedia. After cleaning up the dump, the opening text of each page is extracted as the document[3], with only those longer than 256 characters kept in the annotation pool.

Annotators. Given the complexity of the task, this study recruits native Japanese speakers with expertise in annotating language resources instead of crowdsourcing[4]. The annotators first work individually on different data and then cross-check

[3]2023-01-01 dump at `https://dumps.wikimedia.org/jawiki/`

[4]Measures including the Inter Annotator Agreements (IAA) are thus not reported in this study.

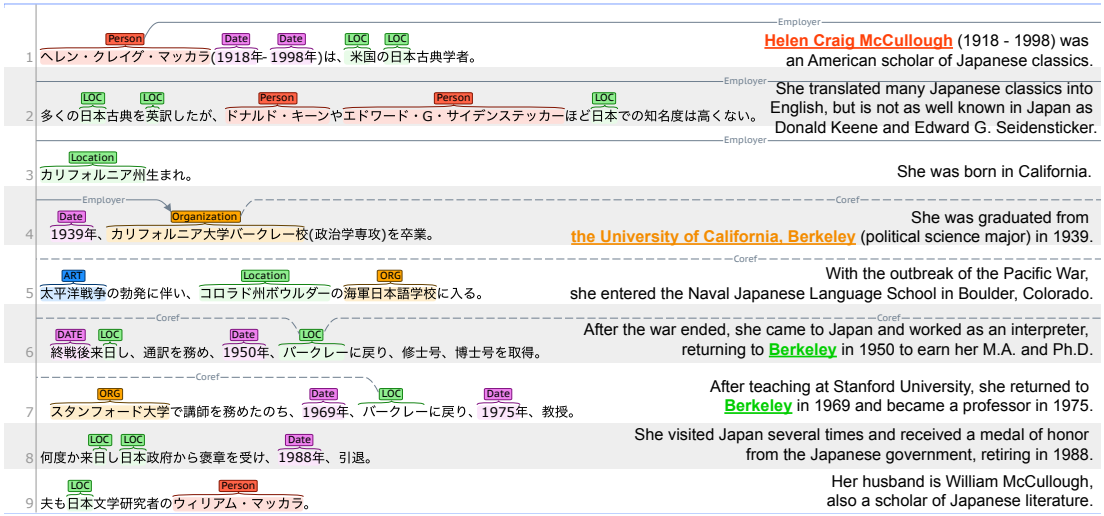


Figure 5.5: Interface for relation annotation. English translations are provided on the right for reference. In this example, the annotator decides whether (Helen Craig McCullough, Employer, the University of California, Berkeley) holds or not. Entity mentions connected with *Coref* are coreferences of each other.

the worked annotations.

Annotation Period. The first phase of entity annotation took 1.5 months to annotate 2,800 documents, and the second phase of relation annotation took 3 months to annotate 2,000 documents in total.

Annotation Interface. The annotation tool is BRAT [87] during both phases, with an example during the relation annotation phase shown in Figure 5.5.

During the **entity annotation phase**, entity mentions recommended by the model were already included as span annotations. In other words, the initial interface window provided to the annotators resembles Figure 5.5, but without any relation annotations (i.e., no arcs between spans). The annotators review and revise all annotated spans in the document, then augment for new spans missed in machine recommendations.

During the **relation annotation phase**, two kinds of interfaces are provided for different purposes. Firstly, to revise the machine recommendations, only *one*

relation triple is provided in one window, as shown in Figure 5.5. The annotators focus on the provided relation triple, decide if it is correct, and supply evidence for the triple. Displaying only one relation instance per window ensures that the target relation triple is clearly visible to the annotators, preventing their attention from being diverted to other triples.

After all machine-recommended triples are revised, a new window displaying all revised relation triples is presented to the annotators. They then review the document again to supplement any missing relation triples in the merged window. This merged window provides a comprehensive view of all annotations, eliminating the need to check each window individually to see if a triple is already included in the machine recommendation, thus saving time.

Additionally, another window displaying all coreference relations is presented to the annotators before revising relation annotations. This allows the annotators to revise the coreference recommendations in advance.

Entity Annotation

The purpose of the entity annotation phase is two-fold: (1) to obtain high-quality entity mention annotations for each document and (2) to filter out documents involved with few entities and relations.

Entity Types. The definition of entities adopts that of IREX (Information Retrieval and Extraction Exercise, Sekine and Isahara [85]). 8 types of entities are included in the annotation scheme, whose scope is similar to that of DocRED. Table 5.2 provides a list of entity types.

Machine Recommendations. The machine predictions of each document are obtained using KWJA [92], a unified analyzer for Japanese.

Document Filtering. Another round of document filtering is performed based on the machine prediction to remove documents likely to contain few **cross-sentence** relations. To this end, each entity mention automatically recognized by KWJA is linked to Wikidata entities [96]. If an edge with label r connects a certain entity pair (e_s, e_o) in the knowledge base, then (e_s, r, e_o) is regarded as an extractable relation triple from the document, following the distant-supervision assumption [71]. Notably, this study preserves only documents with

(Re-)DocRED (6)	JacRED (8)
PERSON	PERSON
ORGANIZATION	ORGANIZATION
LOCATION	LOCATION
TIME	ARTIFACT
NUM	TIME
MISC	DATE
	PERCENT
	MONEY

Table 5.2: Comparison of entity types of existing dataset and our proposed dataset. The total number of entity types is indicated in the parenthesis following each dataset.

more than 4 cross-sentence relations in the annotation pool and eliminates others. mGENRE [26] is employed for entity linking and KGTK [46] for connectivity check in the KB.

Human Edits. 2,800 documents are randomly selected from the annotation pool for human annotation[5]. Human annotators review recommendations in each document, correcting wrongly predicted entity mentions and supplementing missed ones.

Relation Annotation

Relations and coreferences are annotated based on entities. The proposed approach differs from existing studies in that:

1. The relation label set is refined with a reduced number of relation types while covering a sufficient number of relation instances;

2. Machine recommendations are provided by models trained on Re-DocRED$^{\text{ja}}$.

Coreference Recommendations. For each entity e_i, all its mentions $\{m_1^i, \ldots, m_l^i\}$ are treated as coreferences of each other. As introduced in the task definition, only

[5]The final JacRED with both entity and relation annotations contains the first 2,000 among the 2,800 documents (Table 5.1).

proper nouns are considered as mentions while excluding the pronouns. Mentions linked to the same Wikidata entity are recommended as coreferences.

Relation Types. (Re-)DocRED's relation label set $\mathcal{R}$ contains 96 relation types, excluding the null label *no_relation*. However, it is hard for annotators to comprehend such a large label set, which will eventually affect the annotation quality. This study thus reduces the relation label set for human annotation based on the following principles:

- All relation categories defined in ERE [86] should be covered.
- Explicitly-defined inverse relation pairs, e.g., *has_part* and *part_of*, are merged into one. The inversed relation labels are supplemented automatically after the human annotation phase.
- Relations frequently appearing in Re-DocRED are preserved as much as possible.

This results in a label set $\mathcal{R}'$ of 28 relations covering over 88% relation instances in Re-DocRED.

Relation Recommendations. To recommend relation instances, this study projects the relation label set of Re-DocRED$^{\text{ja}}$ from $\mathcal{R}$ to $\mathcal{R}'$ and retrain a DocRE model. Predictions of the model are employed as machine-recommended relations. These recommendations are expected to be more accurate than those in previous studies obtained from knowledge base queries, primarily due to two factors:

- The pre-defined KB, i.e., Wikidata, only stores a limited number of relation facts, while a model can, in principle, assign relation(s) to each entity pair in the document;
- Relation facts in Wikidata are independent of the document's content, while model predictions are contextually sensitive.

Quantitative comparisons of recommendations from the model trained on Re-DocRED$^{\text{ja}}$ and those from querying Wikidata can be found in Section 5.2.2.

Human Edits. Coreferences and relations are revised during human annotation. For coreferences, human annotators remove irrelevant mentions and supplement missed mentions for each entity. For relations, human annotators first examine the existence of each recommended relation. As showcased in Figure 5.5, a pair of mentions m_i^s, m_i^o, representing entity e_s, e_o respectively, along with their relation r is displayed in the interface. If annotators consider relation triple (e_s, r, e_o) as true, they need to provide the evidence sentence $\mathcal{V}_{e_s,r,e_o}$ within the document[6]; Otherwise, the triple should be deleted from the dataset. After revising all machine recommendations, the annotators supply missing relation triples and evidence sentences with their best effort.

This study delicately designs a guideline for relation annotations, as shown in the Appendix. Notably, while in the guideline of existing studies, verifying existing annotations and supplementing missing annotations are listed as parallel tasks, this study explicitly separates the tasks into two parts. The proposed guideline explicitly sets goals and interfaces for (1) verifying existing annotations and (2) supplementing missing annotations, respectively. Annotators are forced to spend time on both (1) and (2), alleviating the dilemma of choosing only to perform (1) rather than (2) as described in Huang et al. [44]. Such a design, in combination with the recruitment of experts instead of crowdsourcing workers, is expected to improve the quality of the constructed dataset.

Post-processing. Among all 28 relation label types, 7 have inverses defined in Wikidata. This study automatically augments triples of inversed relation types after human edits. For example, if triple $(e_s, part\_of, e_o)$ is present in the revised annotation and relation type $part\_of$ is an inversion of has_part, a new triple (e_o, has_part, e_s) will be automatically added into the annotation. JacRED thus includes 35 relation types eventually. Table 5.3 provides a detailed list of relation types.

[6]Sentences where mention m_i^s and m_i^o reside are treated as evidence by default. Only evidence sentences other than those need to be provided.

ERE Category	JacRED Type	ID
Physical	Capital	P36
	CapitalOf	P1376
	AdministrativeLocation	P131
	Location	P276
	WorkLocation	P937
General Affiliation	CountryOfCitizenship	P27
	DateOfBirth	P569
	DateOfDeath	P570
	PlaceOfBirth	P19
	PlaceOfDeath	P20
	Follows	P155
	FollowedBy	P156
Personal-Social	Child	P40
	Sibling	P3373
	Spouse	P26
	ParticipantIn	P1344
	Participant	P710
Part-Whole	MemberOf	P463
	HasPart	P527
	PartsOf	P361
Organization Affiliation	HeadOfGovernment	P6
	OwnedBy	P127
	OwnerOf	P1830
	FoundedBy	P112
	Employer	P108
	Operator	P137
	ItemOperated	P121
	EducatedAt	P69
Others (*)	AwardReceived	P166
	Creator	P170
	Performer	P175
	Published	P123
	PresentInWork	P1441
	Characters	P674
	Platform	P400

Table 5.3: Relation types included in our proposed dataset. Column **ID** shows the Wikidata property ID linked to each relation type. The last category **Others** includes relation types undefined in ERE type.

	DocRED	Re-DocRED	JacRED
# Sentences	7.98	7.98	8.39
# Entities	19.51	19.45	17.87
# Relations	12.45	29.77	21.12
# Evidences	1.60	0.88	1.67

Table 5.4: Comparison of (Re-)DocRED and JacRED. Values are averages per document.

5.2 Dataset Analysis

This section reports the analysis results of JacRED to provide a deeper understanding of the constructed dataset.

Firstly, a comparison of the statistics of JacRED against (Re-)DocRED is conducted (Section 5.2.1). The comparison suggests that JacRED combines the advantages of DocRED and Re-DocRED with plenty of relation and evidence annotations. Notably, JacRED contains more relation triples that require cross-sentence information to extract than Re-DocRED, better aligning with the objective of DocRE. Next, this section reports the number of human annotators' edit steps before reaching the final annotations (Section 5.2.2). Here, another comparison has been made to calculate the number of steps starting from recommendations made by knowledge base queries, i.e., the method used in previous studies, and machine predictions, i.e., the proposed method. From the results, significantly more edit steps would be necessary if the human annotation started from machine recommendations suggested by knowledge base queries. The observation indicates that the proposed annotation scheme effectively reduces human efforts.

5.2.1 Detailed Statistics

Table 5.4 details the agreements and differences between (Re-)DocRED and JacRED. In short, JacRED is a well-balanced dataset for both relation extraction and evidence retrieval, comprising more relation instances than DocRED and more evidence instances than Re-DocRED.

	# Evidence Sentences (%)						
	0	1	2	3	4	5	6
# Triples (DocRED)	3.79	49.71	34.49	8.30	2.52	0.82	0.25
# Triples (JacRED)	0.00	33.09	66.60	0.27	0.04	0.00	0.00

Table 5.5: Evidence distribution of DocRED and JacRED. The distribution of DocRED is computed from the training and development set, and that of JacRED is computed from the whole dataset. Long-tail values with a frequency lower than 0.00% are left out.

Document Complexity. As for document length, JacRED shares a similar scale with (Re-)DocRED at both token and sentence levels. As for the number of relation instances, documents in JacRED contain more relation instances than DocRED on average. This implies that the false negative issue is mitigated in JacRED compared to DocRED.

Evidence Annotation. Re-DocRED revises DocRED to alleviate the false negative issue by supplying missed relation instances. However, evidence sentences for those supplied instances are not included in Re-DocRED. In contrast, this study collects human-annotated evidence sentences during the relation annotation phase. JacRED thus better portrays the correlation between relation and evidence sentences than Re-DocRED.

Evidence Count Distribution. Table 5.5 reports the results of an analysis similar to Section 4.3.7 about the evidence distribution, revealing how many evidence sentences are assigned to each relation triple by human annotators. Compared to DocRED, the distribution of JacRED concentrates more on 1 or 2 evidence sentences. On the other hand, while 44.71% of the relation triples in DocRED is assigned with a single evidence sentence, the percentage of single-evidence relation triples in JacRED is controlled to 33.09%. Alternatively, 66.91% of the relation triples in JacRED have multiple evidence sentences, indicating that most relation triples span across sentences. This suggests that the effort to filter out documents with fewer than 4 cross-sentence relations, as automatically assigned by querying Wikidata, successfully retained documents likely to contain many cross-sentence relations (Section 5.1.3).

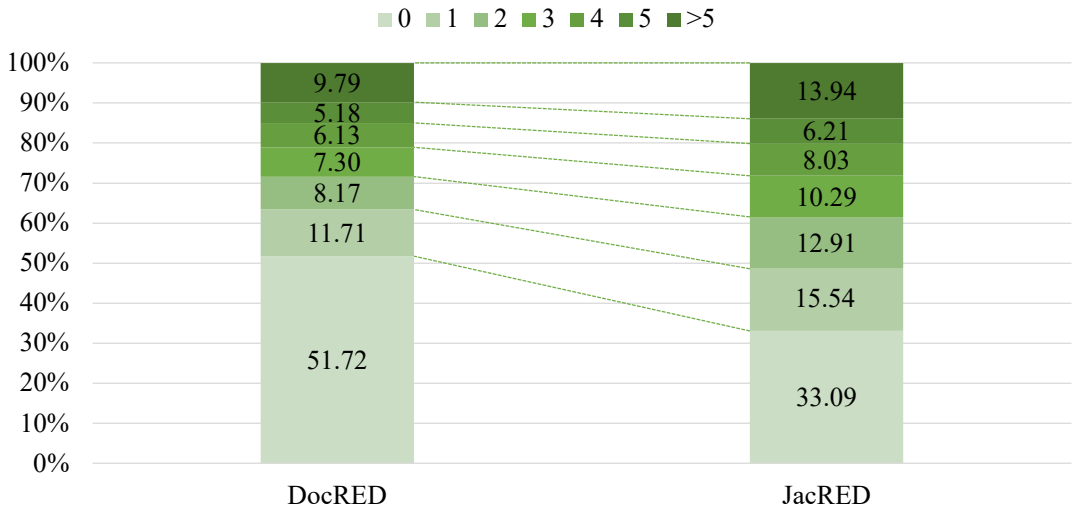


Figure 5.6: Distribution of the distance between evidence sentences in DocRED and JacRED. Distance=0 means that there is only one evidence sentence, distance=1 means the most distant evidence sentence pair is next to each other, etc.

Another important finding behind the statistics is that, although DocRED is claimed to be a document-level relation extraction dataset, approximately half of its relation triples can be extracted from a single sentence. In contrast, JacRED, the dataset constructed in this study, contains more relation instances beyond sentence boundaries in percentage. Therefore, JacRED is more aligned with the objective set by DocRE, focusing on cross-sentence relation extractions.

Distance Among Evidence Sentences. As mentioned in Chapter 1, the focus of DocRE is to extract relation instances beyond sentence boundaries. In response, this study conducts a statistical analysis to see how far evidence sentences spread across the document for each relation instance. Specifically, for an annotated relation instance $(e_s, r, e_o, \mathcal{V}_{s,r,o})$, the maximum distance between any evidence sentence pair $x_1, x_2 \in \mathcal{V}_{s,r,o}$ is recorded, yielding a distribution as in Figure 5.6. For example, if an evidence sentence set is noted as [1,4,5] composing the first, fourth, and fifth sentences within a document, then the distance among evidence sentences is recorded as 5-1=4. Notably, relation instances with

	# Recom.	# Del.	# Sub.	# Supp.
Knowledge Base Queries	3,200	1,459	113	6,233
Model Predictions	6,500	1,266	224	2,740

Table 5.6: Number of relation instances automatically recommended and how they should be revised to reach the final human annotations. *Recom.*, *Del.*, *Sub.*, and *Supp.* are short for *Recommendations*, *Deletions*, *Substitutions* and *Supplements*, respectively.

no evidence annotations are excluded from the calculation[7].

From Figure 5.6, it is clear that JacRED contains more relation instances with a non-zero evidence distance, i.e., cross-sentence relations, corresponding to the findings in the previous paragraph. Furthermore, the distance between evidence sentences in JacRED is, on average, larger than that in DocRED (2.48 v.s. 1.76), indicating that a more comprehensive understanding of the global information within the document is required to solve JacRED.

5.2.2 Number of Human Edits

This subsection quantifies the distance between machine recommendations and human annotations of relation instances. To this end, machine recommendations are compared against final human annotations to see how many edits have been made. The purposes of conducting such an analysis include:

- To depict the difference between machine recommendations and human annotations;
- To depict the difference between machine recommendations made with different approaches.

The analysis is conducted on 400 documents randomly sampled from JacRED. For each sampled document, machine recommendations are conducted using (1) the knowledge-base-querying approach adopted by previous studies and (2) the

[7]The part of data with distance=0 corresponds to that of # Evidence Sentences=1 in Table 5.5. The values vary because those relation instances with # Evidence Sentences=0 are removed in Figure 5.6.

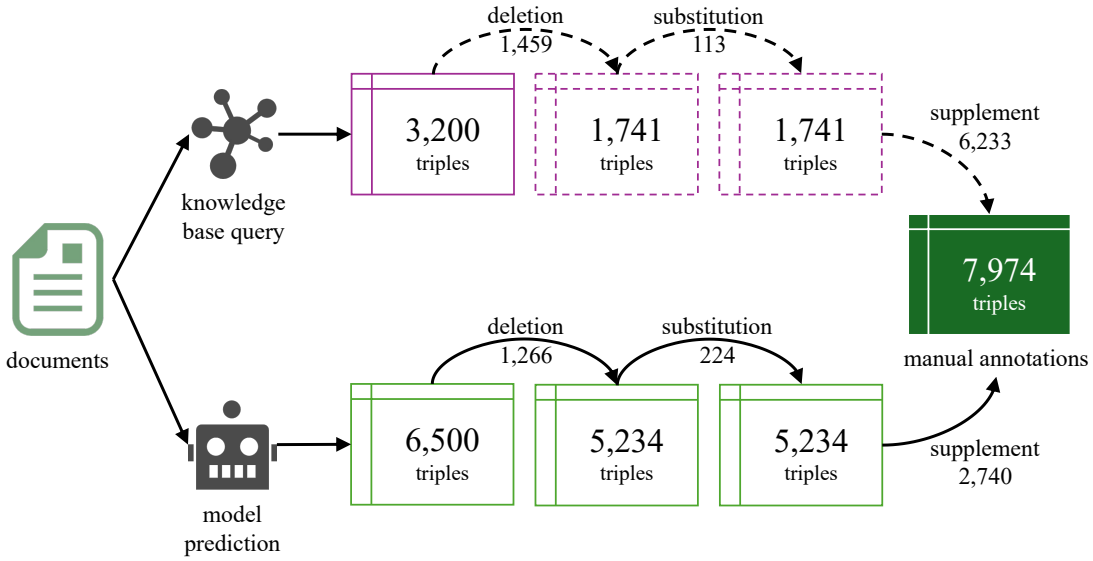


Figure 5.7: Illustration of editing relation instances from different recommendation methods. Recommendations based on knowledge-base queries are simulations drawn with dashed lines.

model-predicting approach proposed by this study. In the annotation pipeline introduced in Section 5.1.3, annotators delete/substitute/supply relation instances based on the recommendations. This section calculates the number of recommended relation triples needed to be be deleted/substituted/supplied to reach the final manual annotations. Notably, human annotations before post-processing are adopted here, i.e., the label set contains 28 relation types with no inverses. The statistics are shown in Table 5.6. Figure 5.7 provides an intuitive illustration that simulates the manual annotation process if it starts with each kind of machine recommendation.

Human Annotations v.s. Machine Recommendations. Table 5.6 shows that even with recommendations provided by model predictions, more than 20% of machine recommendations (1,266 + 224 = 1,490 out of 6,500) were regarded as inappropriate compared to human annotations. Human annotators either deleted these recommended triples or substituted the relation label with another one. The human annotators also supplied another 2,740 relation instances, taking up more than 40% of the recommendations. This, to some extent, explains why DocRE

models trained on the automatically constructed dataset still lag behind human performance considerably, suggesting the importance of a manually constructed dataset.

Knowledge Base Queries v.s. Model Predictions. Table 5.6 also witnesses the distance between human annotations and machine recommendations by querying an existing KB, a de-facto method used in previous studies [18, 112]. Compared with model predictions, knowledge base queries provide only half as many recommendations: To reach the human annotations, 50% (1,459 + 113 = 1,572 out of 3,200) of the recommendations need to be revised, with another 200% instances to be supplied. In total, it takes 7,805 steps to reach the final manual annotations when starting from machine recommendations made with KB queries. In contrast, recommendations made from model predictions require only 4,230 steps to reach the final annotations. Therefore, by training a model on Re-DocRED^{ja}, a dataset automatically constructed from the existing English language resource using cross-lingual projection, the necessary number of human edit steps decreases drastically. These statistics reveal the usefulness of the translated dataset in reducing human annotation efforts.

5.3 Experiments

This section details experiments conducted on JacRED. The major purposes are two-fold:

1. To verify the reasonableness of the proposed annotation method;
2. To employ JacRED as a benchmark and examine the capability of existing DocRE models.

Additionally, the cross-lingual transferability of existing DocRE models is also evaluated by jointly using JacRED and Re-DocRED.

5.3.1 Settings

Dataset. Two datasets are adopted in this section, namely JacRED and Re-DocRED. JacRED is the dataset for Japanese DocRE newly constructed in this

study, while Re-DocRED is the dataset for English DocRE constructed by Tan et al. [89]. Statistics of these datasets are shown in Table 5.1. JacRED is split into train/dev/test sets with 1400/300/300 documents, respectively.

Models. Experiments are conducted on two kinds of models: supervisedly-trained models specialized in DocRE and Large Language Models (LLMs).

For supervised models, the following are selected as the representatives:

- **ATLOP** [124], the baseline model widely used for DocRE;
- **DocuNet** [119], the model that adopts a different strategy than ATLOP to model DocRE as a semantic segmentation task using convolutional networks;
- **KD-DocRE** [88], the model that improves over ATLOP in introducing better representation learning strategies;
- **EIDER** [105], the model that incorporates ATLOP with an evidence classifier, modeling DocRE and ER using different architecture;
- **DREEAM** [70], the model proposed in this dissertation that incorporates ATLOP with evidence-guided attention, modeling both DocRE and ER in the same architecture.

For LLMs, the following models provided by OpenAI are included as the targets of study:

- **gpt-3.5-turbo-instruct** [75], the varient of the GPT-3 model went over an instruction tuning process, during which the model's ability to follow natural language text instructions is improved;
- **gpt-4** [74], the current state-of-the-art LLM showing comparable or even better performance than human beings.

Computation Resources. For supervisedly trained models, all experiments in this chapter adopt PLM encoders with the same scale as $\text{BERT}_{\text{base}}$ [28]. Specifically, for monolingual experiments on JacRED, the PLM encoder is the Japanese

Hyper-parameters	ATLOP	DocuNet	KD-DocRE	DREEAM
# Epoch	30	30	30	30
lr for encoder	5e-5	3e-5	3e-5	5e-5
lr for classifier	5e-5	4e-4	1e-4	1e-4

Table 5.7: Hyper-parameters when training supervised models on JacRED. The PLM encoders are at the same scale as $\text{BERT}_{\text{base}}$.

version of $\text{BERT}_{\text{base}}$ developed by *tohoku-nlp*[8]. For cross-lingual experiments, the PLM encoder is the multilingual version of BERT, noted as mBERT [28]. All models are trained and evaluated on a single Tesla V100 16GB GPU. For LLMs, there is no need to train models on GPU.

Training. For the supervisedly trained models, the training strategy follows what each official repository provides. With a minimal hyper-parameter search, Table 5.7 details the hyper-parameters used in the experiments for each model. For the LLMs, a context-learning strategy is adopted to instruct the model to extract relation triples from a given document (Section 2.2.2, [75]). Details of the prompt are introduced in Section 5.3.5.

Evaluation. The evaluation follows previous studies to compute the micro average F1 scores for relations and evidence sentences [112]. For supervisedly-trained models, average scores of five runs initialized with different random seeds are reported. For LLMs, the result of a single run is reported without computing the standard derivation. Instead of using the whole development/test set, ten documents are randomly sampled to evaluate the performance of LLMs to control the computation cost.

5.3.2 Effectiveness of the Proposed Annotation Method

Motivation. Section 5.1.2 has mentioned limitations in the dataset automatically constructed from cross-lingual projection. Specifically, DocRE models trained on such a dataset fail to extract some relation triples from raw Japanese documents. To resolve the issue, this study recruits human annotators to revise

[8] `https://huggingface.co/tohoku-nlp/bert-base-japanese-v2`

	Relation		
Training Data	**Precision**	**Recall**	**F1**
JacRED (1,400)	$\mathbf{64.76}_{\pm 1.57}$	$\mathbf{73.29}_{\pm 1.49}$	$\mathbf{68.73}_{\pm 0.28}$
Re-DocRED^{ja} (3,053)	$56.14_{\pm 0.56}$	$53.67_{\pm 0.66}$	$54.87_{\pm 0.35}$
Re-DocRED^{ja} (1,400)	$55.52_{\pm 1.26}$	$51.77_{\pm 0.80}$	$53.56_{\pm 0.22}$

Table 5.8: Precision, Recall, and F1 scores of DREEAM trained on different data, evaluated on the test set of JacRED. The number of documents in each set is shown in parentheses.

automatic annotations from models trained on the translated dataset, assuming that results in a dataset of higher quality. This section aims at verifying the effectiveness of introducing such a human revision process. If the quality of the dataset revised by human annotators surpasses the automatically constructed one, then the proposed annotation method is effective.

Results. The experiment results when adopting the test set of JacRED as the benchmark are summarized in Table 5.8. As the language resource is constructed with the intention of training models with expertise in DocRE, the experiments are conducted on supervised models only, using DREEAM as a representative.

Models trained on the translated dataset suffer from low recalls. From Table 5.8, DREEAM trained on Re-DocRED^{ja} underperforms its equivalent trained on JacRED. Taking a closer look at the scores, the gap in recalls (73.29 v.s. 53.67) is more significant than that in precisions (64.76 v.s. 56.14). On the one hand, the results correspond to the observation in Section 5.1.2 that models trained on the automatically constructed dataset cannot identify some relation instances due to the limitation of texts translated from English. On the other hand, the result that models trained on JacRED outperform those trained on Re-DocRED^{ja} verifies the superiority of the human-revised dataset.

The gap between models trained on the translated dataset and JacRED is evident under the same setting. DREEAM is also trained on Re-DocRED^{ja} with only 1,400 documents, aligned with the number of documents in JacRED. The F1 score drops from 54.87 to 53.56, lagging behind the model trained on JacRED with a gap of 15 F1 points. The results demonstrate that JacRED provides better supervision than Re-DocRED^{ja} with a controlled amount

	Dev Set		Test Set	
	Relation F1	**Evidence F1**	**Relation F1**	**Evidence F1**
ATLOP	$66.53_{\pm 0.32}$	–	$68.04_{\pm 0.15}$	–
DocuNet	$66.67_{\pm 0.25}$	–	$67.66_{\pm 0.32}$	–
KD-DocRE	$67.12_{\pm 0.20}$	–	$68.29_{\pm 0.57}$	–
EIDER	$\mathbf{67.52}_{\pm 0.46}$	$57.54_{\pm 0.66}$	$68.61_{\pm 0.27}$	$57.16_{\pm 0.85}$
DREEAM	$67.34_{\pm 0.18}$	$\mathbf{61.52}_{\pm 0.42}$	$\mathbf{68.73}_{\pm 0.28}$	$\mathbf{62.10}_{\pm 0.21}$
gpt-3.5	13.46	–	13.17	–
gpt-4	24.17	–	27.45	–

Table 5.9: Models' performance on the development and test sets of JacRED, with best scores **bolded**. Performance of *GPT-3.5* and *GPT-4* is measured on a single run, and no standard derivation is reported.

of training data, revealing the improved quality of the dataset constructed using the proposed approach.

5.3.3 JacRED as a Benchmark

Motivation. Having verified the properness of the proposed annotation scheme, this study moves forward to utilize the constructed dataset in assessing models' ability to extract relation triples from documents. To this end, JacRED, the dataset constructed following the proposed annotation guideline, is employed as the benchmark. In addition to supervisedly-trained models, the performance of LLMs is also evaluated. Existing studies have observed a large performance gap between supervisedly-trained models and LLMs on Re-DocRED [60]. As LLMs are becoming an essential component of modern NLP research, it is important to assess if the same phenomenon can be observed on JacRED.

Results. Table 5.9 summarizes the performance of each model on JacRED. Supervised models are trained using the training split of JacRED, and LLMs are prompted with in-context learning containing 7 examples from the training split.

JacRED introduces extra challenges beyond those in Re-DocRED. In Table 5.9, all supervised models score above 60 on Relation F1. Although acceptable, the performance of each model is worse than their equivalents trained

on Re-DocRED, with a gap of 10 F1 points (cf. Table 5.10). The result suggests potential challenges in JacRED that are absent from Re-DocRED, possibly due to the characteristics of the Japanese language, such as the omission of subjects. Addressing such characteristics may be essential to tackle Japanese DocRE better.

In-context learning of LLMs on JacRED is non-trivial. Apart from models specially designed for DocRE, the performance of LLMs is also assessed using in-context learning. However, as shown in the Table, GPT-3.5 exhibited much lower performance than the DocRE models. GPT-4 improved over GPT-3.5 but still lagged behind the supervised DocRE models. Similar insights have been provided by Wadhwa et al. [98], where in-context learning of DocRE could not be conducted due to the length restriction of the prompt. This study designs a prompt that successfully instructed LLM to conduct DocRE, but the performance is limited. The experiment results thus highlight the challenge of DocRE as a task that LLMs cannot easily tackle.

DREEAM still scores highest on both DocRE and ER. Among all models, DREEAM still exhibited the highest performance on the test set for both DocRE and ER. Particularly for evidence retrieval, DREEAM outperforms EIDER by approximately 4 F1 points, highlighting the model's superiority.

5.3.4 Crosslingual DocRE

Motivation. Although DocRE datasets have been constructed in Chinese [18] and Korean [111], they lay in different domains than (Re-)DocRED. In contrast, JacRED is constructed from Wikipedia following a pipeline similar to DocRED. The domain and label sets of JacRED and (Re-)DocRED thus match each other, enabling the evaluation of cross-lingual DocRE. This study takes the first attempt to measure the cross-lingual transferability of existing models. Using Re-DocRED and JacRED, it is possible to investigate how knowledge learned in one language can help solve DocRE in another language for each model.

Results. Table 5.10 summarizes the evaluation results. Here, models are trained on the training set in one language and evaluated on the test set in another. The relation label set of Re-DocRED is projected onto JacRED using the same method as in Section 5.1.3. Multilingual BERT (mBERT, Devlin et al. [28]) is adopted as the PLM encoder to ensure the multilingualism of trained models.

	Rel (*tgt*)			Rel (*src*)
Model	**Precision**	**Recall**	**F1**	**F1**
(a) *en.* → *ja.*				
ATLOP	**60.59**$_{\pm 2.17}$	31.91$_{\pm 1.52}$	41.76$_{\pm 0.91}$	74.82$_{\pm 0.26}$
DocuNet	60.44$_{\pm 0.69}$	34.50$_{\pm 0.60}$	43.92$_{\pm 0.36}$	75.02$_{\pm 0.24}$
KD-DocRE	58.83$_{\pm 1.50}$	**36.67**$_{\pm 1.83}$	45.14$_{\pm 1.25}$	75.72$_{\pm 0.46}$
DREEAM	60.07$_{\pm 1.04}$	36.36$_{\pm 1.20}$	**45.29**$_{\pm 0.90}$	**77.22**$_{\pm 0.28}$
(b) *ja.* → *en.*				
ATLOP	53.13$_{\pm 2.08}$	48.70$_{\pm 2.98}$	50.72$_{\pm 0.78}$	64.25$_{\pm 0.63}$
DocuNet	52.69$_{\pm 0.49}$	45.85$_{\pm 0.57}$	49.03$_{\pm 0.30}$	64.64$_{\pm 0.11}$
KD-DocRE	**54.22**$_{\pm 0.61}$	50.12$_{\pm 0.82}$	52.09$_{\pm 0.56}$	65.42$_{\pm 0.68}$
DREEAM	51.88$_{\pm 1.04}$	**53.05**$_{\pm 1.44}$	**52.45**$_{\pm 0.72}$	**65.90**$_{\pm 0.33}$

Table 5.10: Cross-lingual performance on the test set of JacRED (*ja.*) and Re-DocRED (*en.*) of models with mBERT as the PLM encoder.

Cross-lingual performance of existing models is limited. From the table, all models exhibited a decreased accuracy in the target language. Unlike sentence-level tasks, DocRE requires not only an understanding of individual sentences but also inter-sentence semantics within the whole document, which improves the difficulty of building cross-lingual models. This may offer a potential explanation as to why translation-based cross-lingual transfer is ineffective for DocRE, despite its successful application in sentence-level RE and OpenIE [40, 52]. The experiments shed light on the unique challenges of cross-lingual DocRE, leaving the topic as a future direction for research.

Models trained on JacRED and evaluated on Re-DocRED yield better results compared to the reverse scenario. While a performance drop can be observed in both directions, the performance drop of *ja.* → *en.* (65.90 v.s. 52.45) is smaller than that of *en.* → *ja.* (77.22 v.s. 45.29). This may also suggest that JacRED is a high-quality dataset, as models trained on it can effectively extract relation triples from documents, regardless of the language differences between training and inference time.

5.3.5 Prompts used for In-Context Learning

This section details how the prompt used in Section 5.3.3 is designed to ensure reproducibility.

Motivation. In previous studies where LLM are utilized for relation extraction [60, 98], the prompt has been designed to return all relation triples within a document. However, it is hard to identify all relation triples across a document at once. Furthermore, most supervised approaches tackle DocRE by classifying relation types entity-pair wise [70, 88, 105, 124]. Therefore, previous prompt designs result in an unfair comparison between supervised models and LLMs, where the latter solves a more difficult task than the former. This study develops a fairer prompt when instructing LLMs to conduct DocRE.

Prompt. The prompt used for the in-context learning of LLM is shown in Figure 5.8. The prompt is composed of three parts:

- A system instruction specifying the target for prediction written by human experts;
- A definition of the relation type to be extracted. The definition of each relation type is generated by GPT-4. Descriptions on Wikidata are not directly adopted because they contain redundant information for Wikidata editors.
- Seven examples showcasing the documents and relation triples extracted from each. Here, documents refer to partial documents[9] by combining the evidence sentences for the corresponding relation triples to reduce the input length.

Following this design, a prompt is defined for each relation label type. LLMs are asked to extract triples with one relation type in each API call.

Before settling the prompt design, pilot experiments are conducted as in Table 5.11. During these experiments, instead of using the whole development set of JacRED, five documents are randomly sampled to control the computation cost. The temperature of generation is set to 0.1.

[9]Definition of partial documents is mentioned in Section 4.3.3.

	Relation		
	Precision	**Recall**	**F1**
(a) target for each API call			
document-wise	7.31	12.84	9.31
entity-pair-wise	6.50	**50.00**	11.51
relation-wise	**12.29**	14.86	**13.46**
(b) strategy for choosing examples			
random	**8.91**	14.86	**11.14**
shortest	2.82	**17.57**	4.86
longest	7.61	9.46	8.43
in-turn	8.60	12.84	10.30

Table 5.11: Pilot experiments for prompt engineering using *gpt-3.5*, evaluated on five documents randomly sampled from the development set of JacRED. Performance is measured on a single run, and no standard derivation is reported.

Target for each API call. There are three possible ways to conduct relation extraction for each API call.

- **document-wise**: query the LLM to enumerate all relation triples referrable from a given document. This is the setting adopted by Li et al. [60].
- **entity-pair-wise**: query the LLM to answer the relation label between a specific entity pair within a given document. The setting is close to that in supervised training, but it requires calling the API $m \times (m-1)$ times for a document containing m entities.
- **relation-label-wise**: query the LLM to enumerate all triples with a specific relation type within a given document. The setting requires calling the API $|\mathcal{R}|$ times when $\mathcal{R}$ is the relation label set.

As shown in Table 5.11a, the last setting yields the best performance in the pilot experiments. Therefore, the final prompt adopted in Table 5.9 performs relation-label-wise queries, asking the model to list all triples with a particular relation label in a document simultaneously. Compared to the entity-pair-wise strategy exhibiting a high recall, this strategy also controls the cost by reducing the number of API calls.

Strategies of choosing examples. The examples or demonstrations included in the prompt can be chosen using three strategies:

- **random**: Randomly sampling k documents from the training set;
- **shortest**: Sort the training set and select the k-th shortest documents;
- **longest**: Sort the training set and select the k-th longest documents;
- **in-turn**: Sort the training set and select the $\frac{k}{2}$-th longest and the $\frac{k}{2}$-th shortest documents.

Table 5.11b indicates that the first setting outperforms the other three and thus is adopted in the final experiments using the whole dataset.

5.3.6 Influence of Topic Shifts

Motivation. Section 5.1.2 has attributed the limitations of the translated dataset to two factors: the topic shift of contents and the difference in surface form. This section delves into the model trained on the translated dataset to see if the attribution is reasonable.

The focus is the topic shift between the translated and manually constructed datasets. The test set of JacRED is divided into two parts: **local** documents with contents about Japanese figures, organizations, or artifacts (e.g.,明石市立図書館, Akshi Library) and **non-local** documents with contents about western culture (e.g.,オランダ高速鉄道, NS Hispeed). Models trained on the translated dataset, i.e., Re-DocRED^{ja}, and on the manually constructed dataset, i.e., JacRED, are evaluated on each of the splits to see how different the performance can be. If models trained on Re-DocRED^{ja} exhibited higher performance on the **non-local** split than the **local** split, that would support the hypothesis that the topic shift causes the limitations of the translated dataset.

Results. JacRED's test set is manually divided into the **non-local** and the **local** split. The non-local split contains 217 documents, and the local split contains 83. Table 5.12 demonstrates the experiment results.

The results support the hypothesis raised above with the following two evidences. Firstly, **models trained on the manually constructed dataset consistently outperform those trained on the translated dataset on both**

Perform Document-level Relation Extraction task. Given a context and an entity list, identify all entity pairs with relation type {located in the administrative territorial entity} in the context. Note that only a few entity pairs hold relations. Please return entity pairs as {head, tail} and make sure they follow the relation definition:

located in the administrative territorial entity: {head} is located in the administrative territorial entity {tail}.

###

Context: 東京・板橋出身。

Entity List: 東京||板橋

Extracted Entity Pairs: {板橋, 東京}

###

Context: 南都六宗(なんとろくしゅう、なんとりくしゅう)とは、奈良時代、平城京を中心に栄えた日本仏教の6つの宗派の総称。三論宗(さんろんしゅう、中論・十二門論・百論)-華厳宗や真言宗に影響を与えた成実宗(じょうじつしゅう、成実論)-三論宗の付宗(寓宗)法相宗(ほっそうしゅう、唯識)倶舎宗(くしゃしゅう、説一切有部)-法相宗の付宗(寓宗)華厳宗(けごんしゅう、華厳経)律宗(りっしゅう、四分律)-真言律宗等が生まれたなお、奈良時代当時から「南都六宗」と呼ばれていたわけではなく、平安時代以降平安京を中心に栄えた「平安二宗」(天台宗・真言宗)に対する呼び名である。

Entity List: 奈良時代||平安時代||平城京||日本||平安京||平安

Extracted Entity Pairs: {平安京, 日本}

###

(examples)

###

Context: アンソニー世界を駆ける(アンソニーせかいをかける)は、アメリカ合衆国のCNNで放送されているテレビ番組。2013年4月から放送を開始した。エミー賞を4回受賞、また、脚本賞、音響賞、編集賞、撮影賞に11回ノミネートされている。また2013年にはアメリカのテレビ・ラジオ・ウェブサイトの優れた放送作品に贈られるピーボディ賞を受賞した。自ら料理人であり、ノンフィクション「キッチン・コンフィデンシャル」の著者でもあるアンソニー・ボーディンが世界の津々浦々を旅し、あまり知られていない地域の景観、風俗、食材、料理などを紹介する。

Entity List: アメリカ合衆国||アンソニー世界を駆ける||CNN||2013年4月||エミー賞||2013年||ピーボディ賞||キッチン・コンフィデンシャル||アンソニー・ボーディン

Extracted Entity Pairs:

Figure 5.8: An example of the prompt used for the in-context learning of GPT-3.5 and GPT-4.

	local (83)	**non-local** (217)	**Δ**
(a) w/o inference stage fusion			
Re-DocRED^{ja}	$47.91_{\pm 0.46}$	$55.92_{\pm 0.20}$	8.01
JacRED	$64.00_{\pm 0.89}$	$68.99_{\pm 0.31}$	4.99
(b) w/ inference stage fusion			
Re-DocRED^{ja}	$49.80_{\pm 0.42}$	$57.01_{\pm 0.24}$	7.21
JacRED	$64.04_{\pm 0.77}$	$70.06_{\pm 0.35}$	6.02

Table 5.12: Performance of DREEAM on the global and local split of JacRED's test set. The number of documents in each set is shown in parentheses.

splits. Particularly, on the local split containing documents about Japanese culture, there is a significant performance between models trained on Re-DocRED^{ja} and JacRED, regardless of inference stage fusion. The observation suggests that DREEAM trained on JacRED is better at extracting relations from documents containing topics about Japanese culture than those trained on Re-DocRED^{ja}. Secondly, **models trained on the translated dataset perform better on the non-local split than the local split.** Specifically, in Table 5.12a, the score of DREEAM trained on Re-DocRED^{ja} on the non-local split was 55.92, leading that of the local split by 8.01 F1 points. While DREEAM trained on JacRED also performed better on the non-local split, the gap was 4.99 F1 points. The observation indicates that DREEAM trained on Re-DocRED^{ja} is much better at extracting relations from documents containing topics about Western culture.

5.4 Summary

This chapter has proposed an annotation strategy employing model predictions obtained using cross-lingual projection to provide recommendations. To reduce human annotation costs in constructing DocRE datasets, this study explores how to utilize existing English DocRE resources to construct resources for other languages, using Japanese as the representative. Initially, it constructs a dataset through translation-based cross-lingual projection and investigates why such a dataset is not ready for practical use. Nevertheless, models trained on the dataset can replace existing approaches, i.e., querying knowledge bases, to provide recom-

mendations for human annotation. These insights can benefit the development of DocRE resources for other languages.

Using the proposed strategy, this study constructs and publishes JacRED, the first benchmark for general-purpose Japanese DocRE. JacRED has a relation label set smaller than its English counterpart, which comprises a considerable number of relation and evidence instances. Analyses on JacRED suggest that the proposed machine recommendation strategy using model predictions succeeds in reducing human annotation steps to half, compared with existing methods.

Experiment results on JacRED verify that the proposed annotation process improves the dataset quality over the automatically-constructed counterpart. Benchmarking with JacRED portrays the challenge of not only Japanese but also cross-lingual DocRE. Notably, it is difficult to handle DocRE using large language models, even when combined with an in-context learning strategy. This empirical finding highlights the importance of developing strong supervised DocRE models, as it automates the process of knowledge base completion.

6 Conclusion

This dissertation presented two proposals to accelerate information extraction beyond sentence boundaries. Both methods target the same task, Document-level Relation Extraction (DocRE), which aims at extracting all relation tiples from a document composed of multiple sentences. Given the complexity of the task, acquiring high-quality human-annotated data for DocRE is both time-consuming and expensive. This study improves the situation of DocRE from two different aspects: to make full use of existing annotations and to collect new annotations with a reduced cost.

The first proposal utilizes all available human annotation data to **train a better DocRE model**. To this end, the proposal puts a special focus on the evidence annotation provided in DocRED [112], the first and most popular general-purpose DocRE dataset. Existing studies train an evidence classifier to retrieve the evidence for each relation triple automatically, which is designed separately from the relation extractor. Such a design limits the usage of evidence supervision signals when updating the parameters in the relation extractor. To address the problem, this study proposes a mechanism, Document-level Relation Extraction with Evidence-guided Attention Mechanism (DREEAM), to merge the model of evidence retrieval into that of relation extraction. The supervisory signals of evidence retrieval, namely the evidence annotations, are injected directly into the entity-pair encoder that computes the contextualized representation used for relation classification. In such a way, the evidence annotation can also be utilized to train the relation classifier. In conjunction with DREEAM, a weakly-supervised training strategy for evidence retrieval is also proposed. The strategy assigns silver (pseudo) evidence to unlabeled data, significantly increasing the amount of data that can be utilized for training relation extraction and evidence retrieval. The superiority of DREEAM over other existing methods has been verified in experiments. Detailed analyses investigate how evidence-guided attention contributes to training a better DocRE model.

The second proposal explores a new scheme to **construct new human-annotated language resources** for DocRE with a reduced cost. Especially, the study focuses on constructing DocRE datasets in non-English languages, choosing Japanese as the representative. Inspired by the success of cross-lingual projection in constructing multilingual sentence-level relation extraction datasets [40], the study explores automatic dataset construction by translating datasets in English to Japanese. However, experiments have revealed the drawbacks of the translated dataset, which is not ready for practical use. The finding highlights the difference between DocRE and sentence-level RE. The study then utilizes the translated dataset to train a DocRE model on the target language to provide relation triple recommendations, which are further revised by human annotators. The recommendation strategy based on model prediction improves over existing studies, where only relation triples present in a pre-defined knowledge base are recommended to annotators. Apart from the initial seeds recommended for human annotation, several tricks, such as reducing the size of the relation label set, are also included in the proposed guideline to reduce the burden of human annotators. The dataset constructed following the proposed annotation scheme, JacRED, is published as the first Japanese Document-level Relation Extraction dataset. Analysis of the dataset demonstrates that the proposed annotation scheme, which begins with model predictions, successfully cut human annotation costs by half. The dataset portrays the task DocRE well, with more than 60% of the relation instances residing beyond sentence boundaries. Notably, while a majority of the cross-lingual dataset construction strategy ends with machine translation [55, 58], this study went further. Instead of stopping after obtaining a translated dataset, it investigated the limitations of the obtained dataset and proposed strategies for improvement. The proposed annotation scheme is theoretically appliable to collecting DocRE datasets in any language.

Proposals in this dissertation complement each other and promote DocRE research together. Specifically, the issue of data scarcity is addressed during the development of DREEAM, elevating the performance of DocRE models in languages where datasets are already constructed. JacRED and how it is constructed expand the use of DocRE to new settings where datasets are not yet available.

Since DocRE, as a representative of information extraction beyond sentence boundaries, is driven by the goal of automating knowledge base completion, this dissertation further contributes to the field of Natural Language Processing (NLP)

by offering methods and insights specifically for enhancing knowledge base completion. With better knowledge base completion methods, it will be easier to organize and manage the unstructured data. The need to structuralize data and organize the information not only resides in data publicly available on the World Wide Web but also in confidential data in administrations or companies. This dissertation, therefore, provides advancements for both the NLP fields and broader societal applications, improving data accessibility and utility.

Finally, the limitations of this work and potential future directions are discussed as follows.

DocRE on long documents. Although the study addresses information extraction beyond sentence boundaries, the average document length examined remains at no more than 300 tokens (Table 5.1). Therefore, models and annotation strategies described in this dissertation cannot be directly adapted to long documents with multiple paragraphs or pages. Extracting relation triples from long documents remains an unsolved challenge. Potential solutions include: (1) increase the processable length of DocRE models: the maximum length processable for a DocRE model is decided by the maximum input length of the pre-trained language model encoder. Therefore, switching the encoder to Longformers could provide a possible solution [9]. (2) reduce the length of documents: splitting the long document into multiple chunks or summarizing the long document using supervised summarization models or LLMs [13, 95].

Improving the annotation method. Although this study presented a new method for constructing DocRE datasets, the method is imperfect. For annotating relations, although being mitigated, it can be foreseen that false-negative issues are still present in JacRED. Another round of human-machine collaborated annotation, similar to that adopted in constructing Re-DocRED [89], could be employed to further reduce the number of false-negatives in relation annotations. For annotating evidence, cases may exist where evidence is over-annotated (i.e., sentences more than necessary are annotated) or under-annotated (i.e., not enough sentences are annotated). To improve the reliability of evidence annotations, a solution is to recruit multiple human annotators to evaluate whether the evidence annotation of the same triple is proper, then measure the inter-annotator agreement rate [4, 11, 54].

Annotating DocRE Corpus in other languages. This study has proposed an annotation scheme assisted by existing language resources, whose effectiveness

has been observed in Japanese. However, to verify the soundness of the proposed scheme thoroughly, it is preferred to construct datasets in multiple languages. Due to the constraints of time and resources, these explorations are deferred to future studies.

Cross-lingual DocRE. As depicted in Section 5.3.4, existing DocRE models have a limited cross-lingual transfer ability. Developing a system that better addresses cross-lingual DocRE is another potential research direction. This effort could also reduce human annotation costs by providing machine recommendations of even higher quality.

Resolving the conflicts of extracted relation triples. To ensure the effectiveness of DocRE models in knowledge graph completion, merely extracting relation triples from documents is not enough. An extra step needs to be conducted to resolve the conflicts between extracted relation triples, and those with existing triples in the knowledge base. Although such a post-processing step is out of the scope of this dissertation, the direction should be considered seriously when applying DocRE to actual use.

Combining DocRE models with LLMs. Nowadays, as LLMs are growing to be the foundation models of modern NLP research, it is necessary to redefine the purpose of DocRE research in the LLM era. As mentioned in recent works, there is an increasing need for knowledge bases to improve the reliability of LLMs by alleviating hallucinations and improving the reasoning abilities [1, 6, 68, 122]. It is thus a promising research direction to put DocRE models into actual use for automatically completing knowledge bases, which can be adopted as an information source for LLMs. Furthermore, LLMs can be utilized to provide supervisory signals for DocRE models, which in turn will enhance their ability for knowledge base completion.

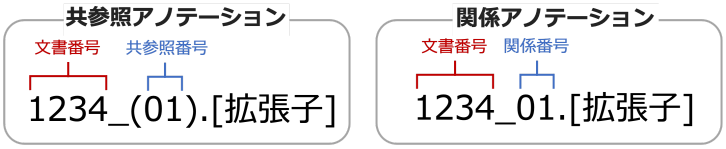
Appendix

This section outlines the guideline used in relation annotation phase described in Chapter 5.

日本語文書レベル関係抽出アノテーションマニュアル

1. 作業内容

日本語のWikipedia文書2,000件に対して下記1.1，1.2の作業を行う．
ファイル名のフォーマットは以下の通り．上四桁を文書番号，下二桁を共参照・関係番号とする．文書番号と共参照・関係番号は，アンダーバー（_）で繋げる．



ただし，同じ文書番号を共有するファイルのテキストは同一であることに注意されたい．

1.1. 共参照アノテーションの修正

提供された文書には，既存の自然言語処理技術で共参照アノテーションが付与されている．共参照関係は，同じ実体を指す固有表現の間にだけ存在する（代名詞を考慮しない）．
なお，一つのアノテーションファイルにつき，共参照関係が1セットだけある．

1. 共参照関係が付与された複数固有表現のうち，実は違う実体を指す固有表現があった場合，該当固有表現を繋ぐ共参照関係を削除する．
2. 既に共参照関係が付与された固有表現のほかにも，同じ実体を指す固有表現があった場合，該当固有表現との共参照関係を追加する（＊既存共参照セットの固有表現添削について）．
3. 自動認識されなかった共参照関係のセットを追加する（＊共参照の追加について）．

1.2. 関係アノテーションの修正及び根拠文の提示

提供された文書には，既存の自然言語処理技術で関係アノテーションが付与されている．一つのアノテーションファイルにつき，関係が1つだけある．
各ファイルの関係は，**テキストの文脈**から推測できるもの（すなわち，専門知識がない人間が文書を読んだ上で，常識に基づいて推論できるもの）か精査する．簡易のため，自動付与された関係ラベルを`rel`とする（＊専門知識と常識）．

1. 関係`rel`が文脈から推測できず，専門知識が必要か，そもそも関係が間違っている場合，関係`rel`を削除する．
2. 関係`rel`が文脈から推測できる場合，推測の根拠を文単位で提示する（＊根拠文の提示について）．
3. 自動抽出されなかったが，文脈から推測できる関係ラベル`rel`を付与する（＊関係の追加について）．

2. 作業ツール

brat

3. 作業段階

アノテーション作業を二段階で行う．

3.1. 第一段階

文書番号0000~0399の文書に対し，1.1と1.2の作業を行う．

納品対象：文書番号0000~0399のまとめファイル（`.txt`及び`.ann`形式）．関係別ファイルは提供しなくても良い

第一段階の結果に基づき，作業依頼者の方で関係抽出器を再学習し，関係の自動付与結果を更新する．

3.2. 第二段階

残り全ての文書（文書番号0400~1999）に対し，1.1と1.2の作業を行う．

納品対象：文書番号0400~1999のまとめファイル（`.txt`及び`.ann`形式）．関係別ファイルは提供しなくても良い

なお，自動アノテーションは更新後のものを使う．

4. 作業手順

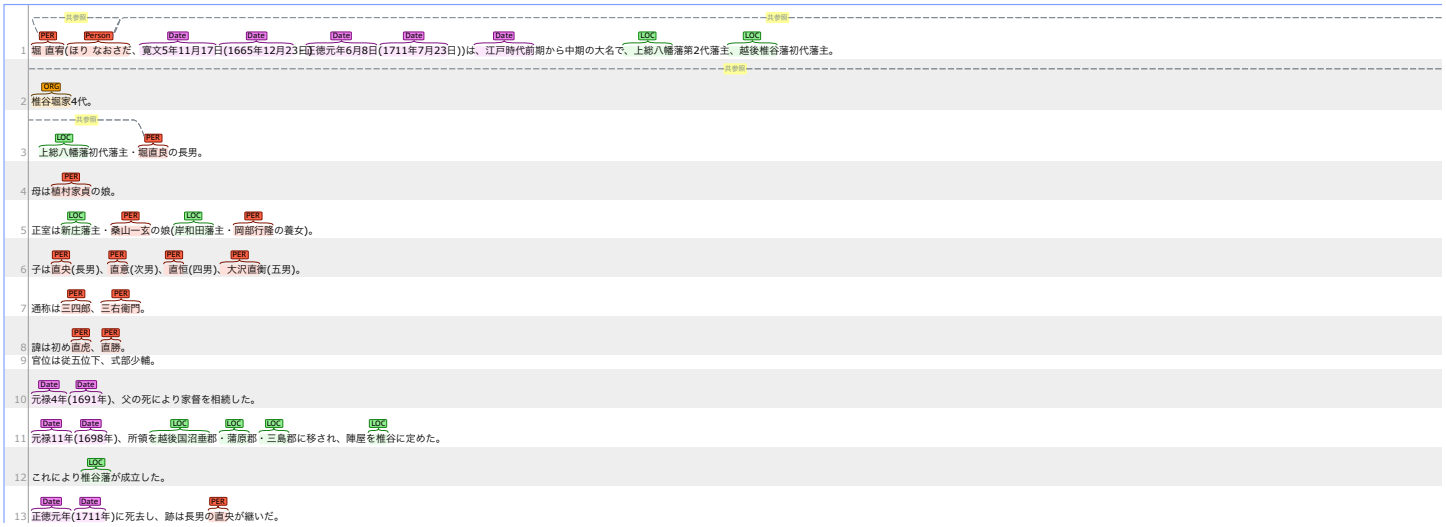
文書0002を例に，アノテーションの手順を具体的に説明する．

4.1. 共参照アノテーション

(1) 自動認識共参照セットの固有表現添削

まず自動認識された共参照のセットが正確であるかを精査する．自動認識された共参照は，セット毎にアノテーションファイルが作られ，おおよそ**文書での出現順**で並べられている．なお，各ファイルにおける添削作業をし終えた次第，次のファイルに進んで良い（ファイル毎で文書全体を読む必要はない）．

例えば，アノテーションファイル`0002_(00)`（文書0002の00番目の共参照セット）の中身は以下とする．



このファイルでは，1番目の文にある固有表現「堀　直宥」と「ほり　なおさだ」・3番目の文にある固有表現「堀直良」が共参照として認識されている．
以下の理由から，当共参照セットは不適切であり，添削が必要である．

- 「堀直良」だけ別の実体を指している
- 「三四郎」・「三右衛門」・「直虎」・「直勝」も「堀　直宥」・「ほり　なおさだ」と同じ実体を指している

添削作業の内容は以下である．

- 誤認識された固有表現（「堀直良」）と他を繋げる関係ラベル「共参照」をダブルクリックし，それを削除する
- 認識されなかった固有表現（「三四郎」・「三右衛門」・「直虎」・「直勝」）を既に共参照関係のある固有表現とそれぞれ繋ぎ，関係ラベル「共参照」を付与する．なお，既に共参照関係のある固有表現であれば，どれに繋げてもよい

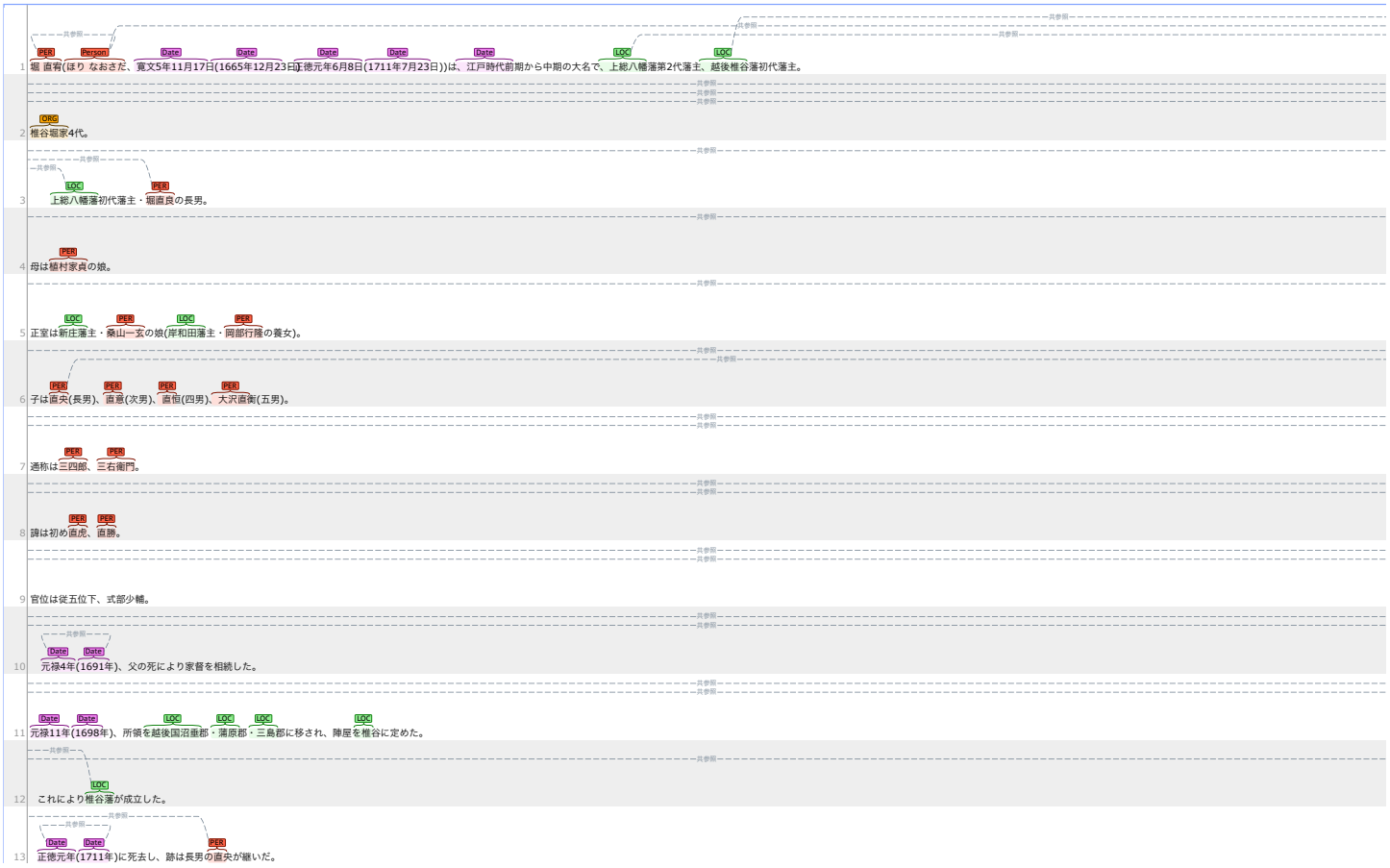
最終的には，同じ実体を指す全ての固有表現だけに共参照関係が付与されていることを確認されたい（以下参照）．

なお，基本的には文字列が同じであれば共参照とし，関係ラベルを付与する際に不整合が生じた場合だけベストエフォートで修正する．

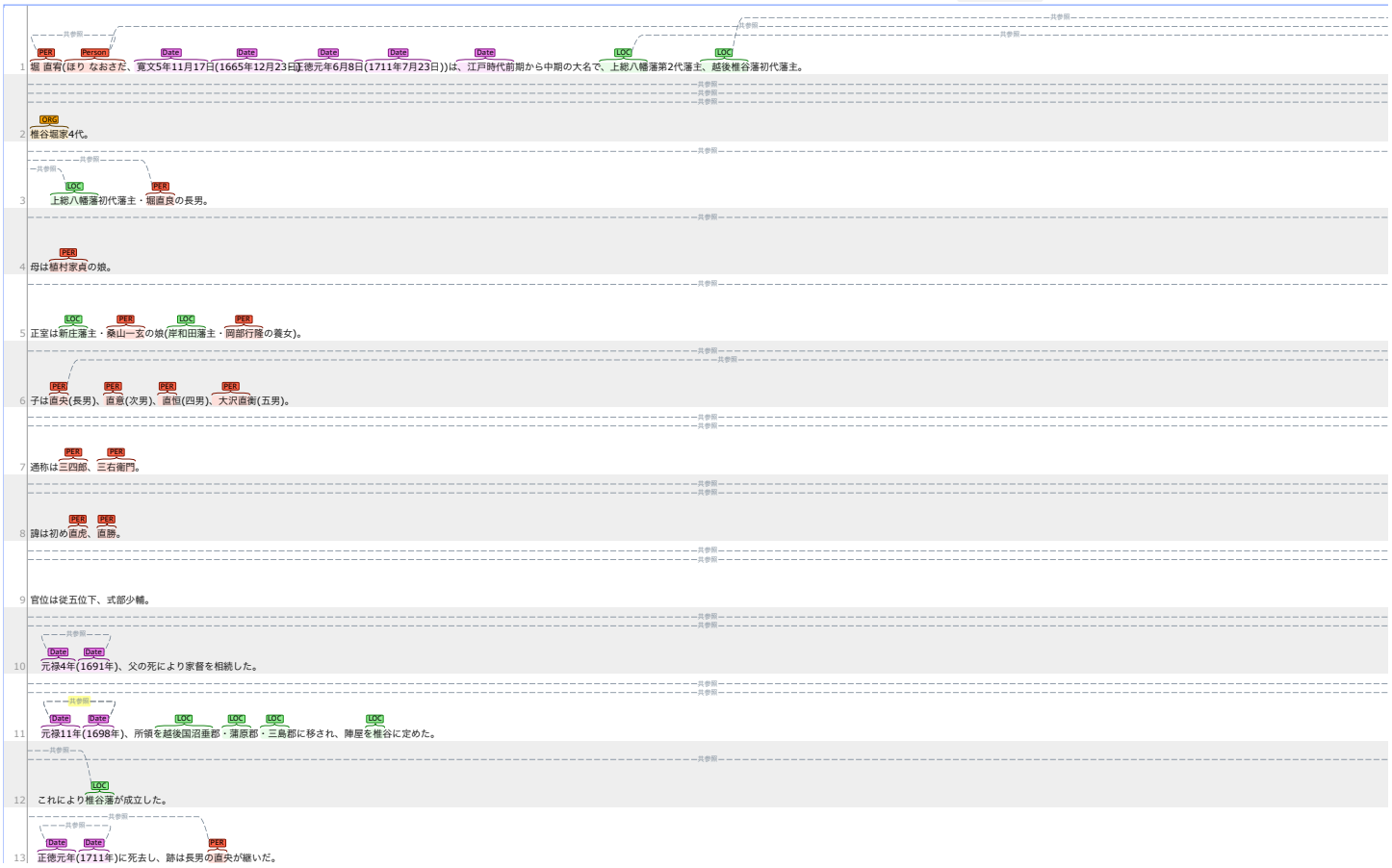
(2) 新規共参照セットの追加

自動認識されなかった共参照セットを全ての共参照アノテーションをまとめたファイルに追加する（まとめファイルは作業者の都合の良いように作成して頂きたい）．なお，全ての共参照セットを検出することを目的とせず，ベストエフォートで行う．

例えば，文書0002の共参照アノテーションまとめファイル `0002_(all)` の中身は以下とする．



このファイルでは，「元禄11年」と「1698年」の共参照関係が欠落している．そのため，「元禄11年」と「1698年」の共参照関係を以下のように `0002_(all)` に追加する．



4.2. 関係アノテーション

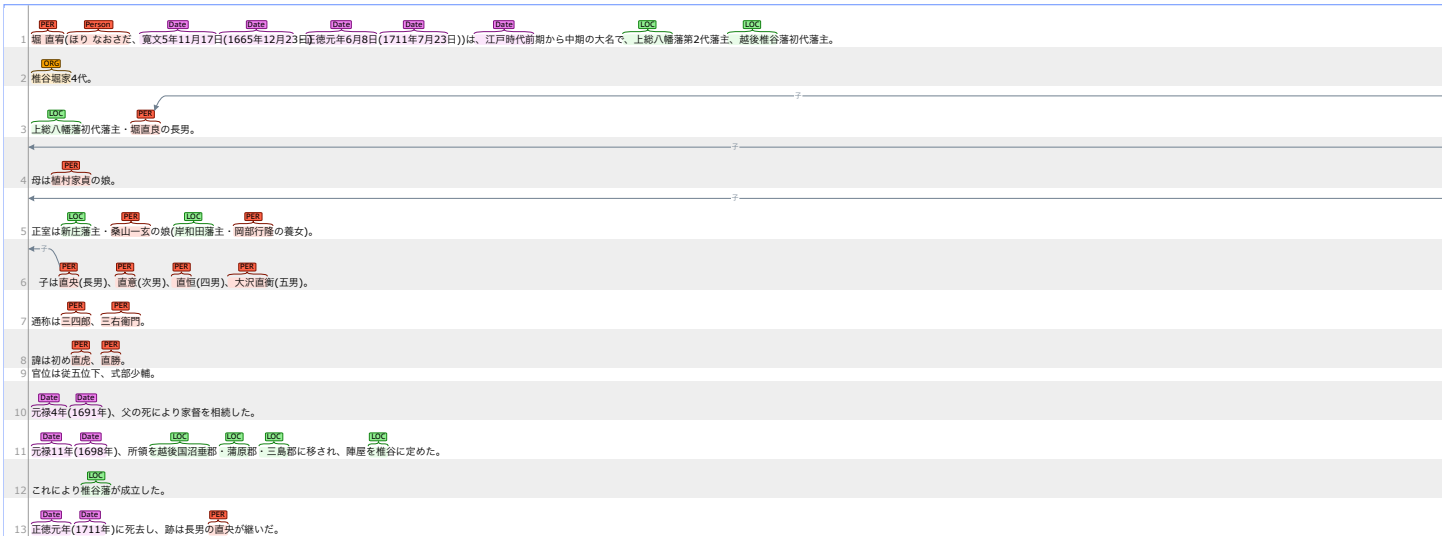
以降では関係事例を（head, rel, tail）の三つ組として表記し，`head` と `tail` は固有表現，`rel` は関係ラベルとする．

(1) 自動認識関係の訂正

まず自動認識された関係が正確かどうかを精査する．自動認識された関係は，**文書での出現順で**ファイル毎に一つだけ配置されている．なお，固有表現のペア一つに対して，関係ラベルが複数存在し得ることに注意されたい．各ファイルにおける作業をし終えた次第，次のファイルに進んで良い（ファイル毎で文書全体を読む必要はない）．

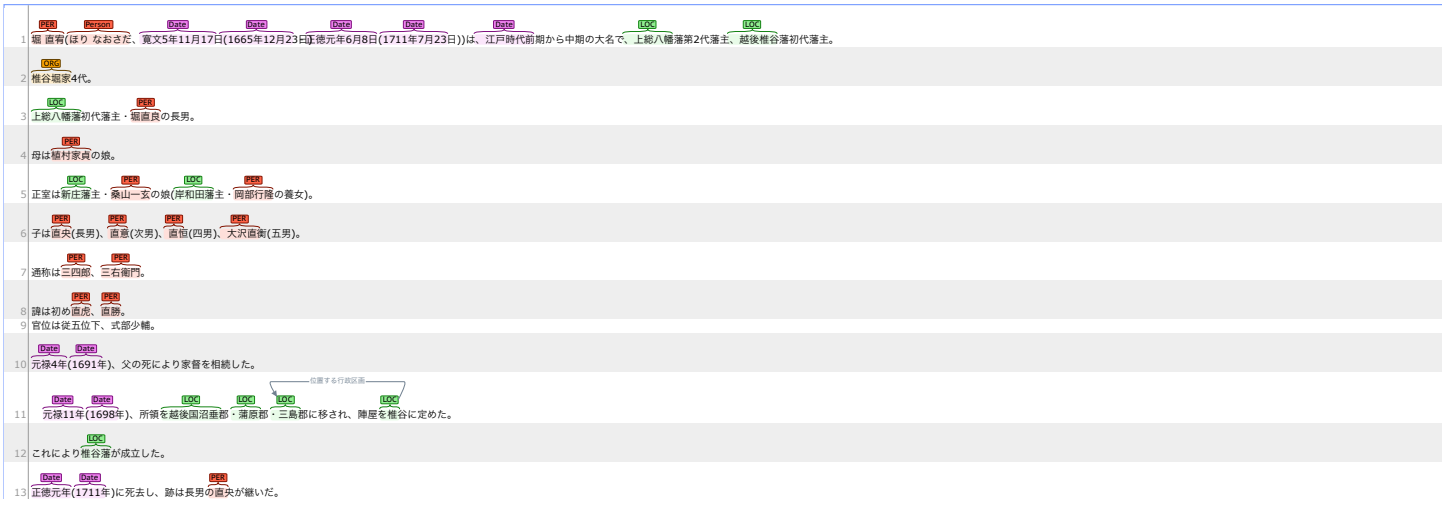
削除すべき関係

文脈から明らかに間違っている関係は削除する．固有表現 `head` と `tail` は文脈的な繋がりがあり，その繋がりから見て関係 `rel` は不適切な場合，関係 `rel` を削除する．
例えば，以下の自動アノテーションを考える．



ここでは，6番目の文にある固有表現「直央」は3番目の文にある固有表現「堀直良」の「子」であると示している．
しかし文脈から見て，「直央」は「堀直良」の子でなく，孫であるため，この関係ラベルが不適切である．
よって，当関係ラベルをダブルクリックし，それを削除する．

文脈から明示されていなく，かつ判断するのに専門知識が必要となる（作業者から見て非自明である）関係は削除する．文脈的な繋がりがなく，かつ常識の枠を超えた関係は，知識としての正しさを問わずに削除する．
例えば，以下の自動アノテーションを考える．



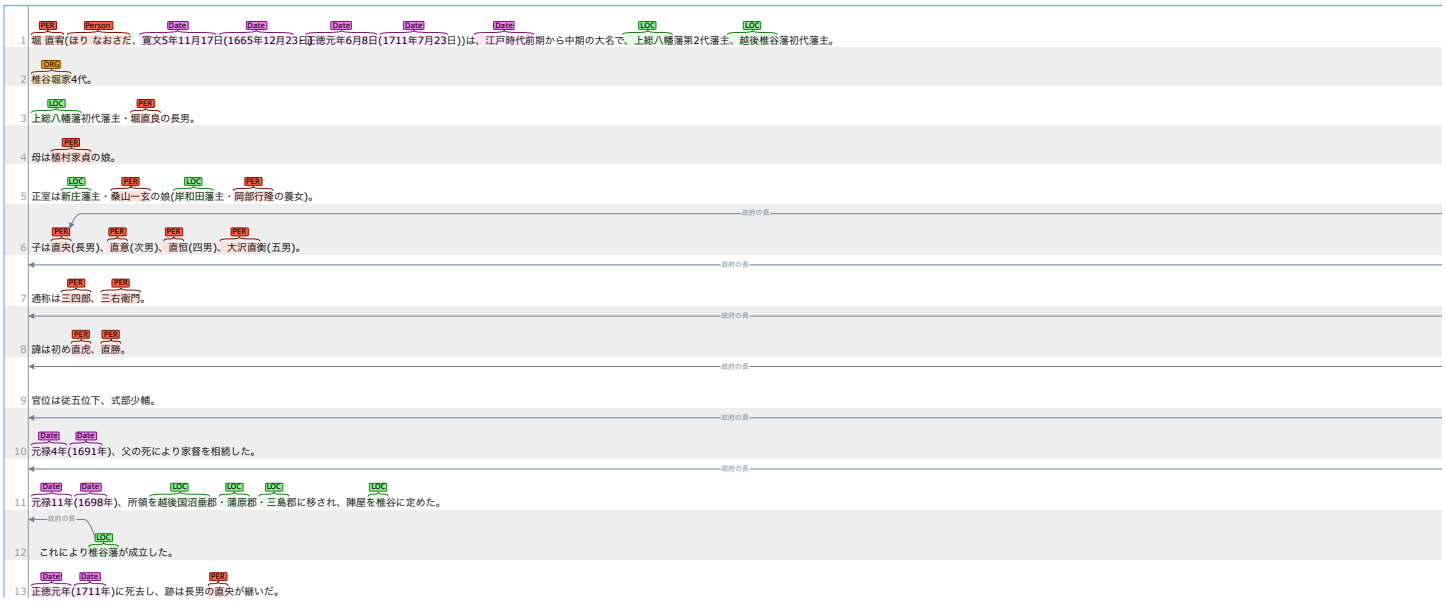
ここでは，11番目の文にある「椎谷」は「三島郡」という行政区画に位置すると示している．この関係は知識的に正しいが，以下の問題点がある．

- 文書から，「椎谷」が「三島郡」にあることを示唆するようなヒントはない
- 常識を加えても，「椎谷」が「越後国沼垂郡」・「蒲原郡」・「三島郡」のいずれかにあるのを推測できるが，「三島郡」までは特定できない

よって，当関係ラベルをダブルクリックし，それを削除する．

残すべき関係

文脈から明示的に示されているものは残す．固有表現 `head` と `tail` は文脈的な繋がりがあり，その繋がりから見て関係 `rel` は適切である場合，関係 `rel` を残す．
一例として，以下の自動アノテーションを考える．

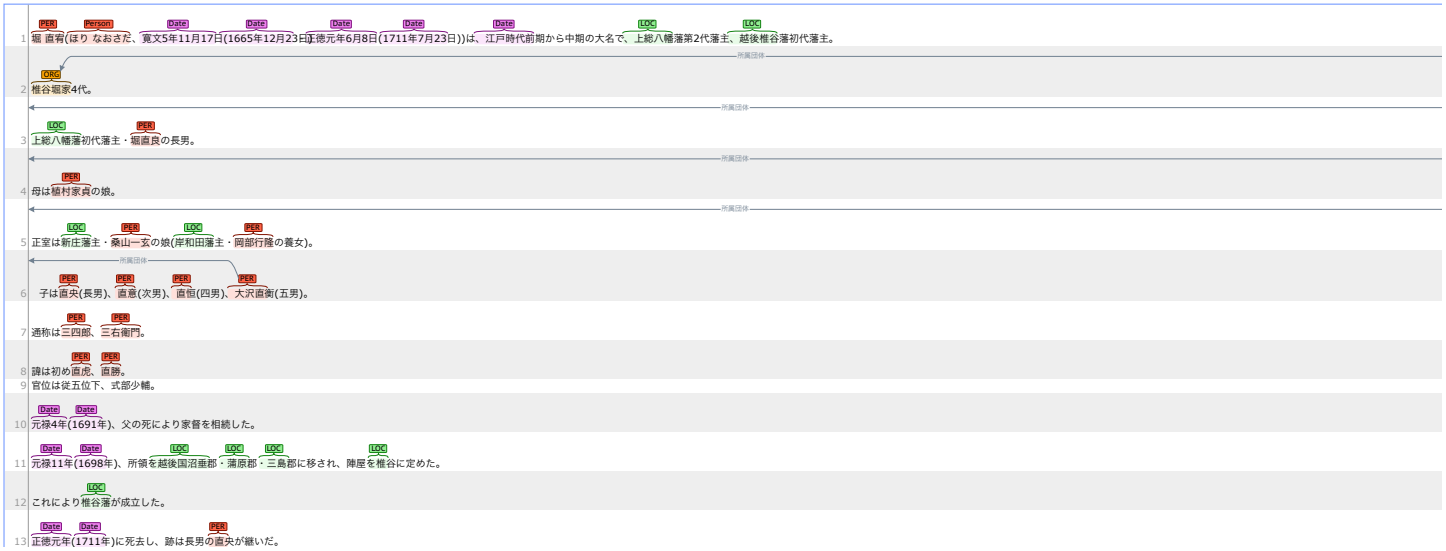


ここでは，6番目の文にある固有表現「直央」が，12番目の文にある固有表現「椎谷藩」の「政府の長」であると示している．
文脈から，以下の情報が汲み取れる．

- 「直央」は「堀　直宥」の長男（文6，文13）
- 「堀　直宥」は「椎谷藩」の藩主（文1）
- 「堀　直宥」の死後，跡を「直央」に継がせた（文13）

よって，関係（椎谷藩，政府の長，直央）が文脈から明示的に示されていることから，正しい関係として残す．

文脈から明示されていないが，常識を加えると推測できるものは残す．明示的に示されていないが，固有表現 `head` と `tail` は文脈的にある程度の繋がりがあり，さらに常識を加えると関係 `rel` が推測できれば，関係 `rel` を正しいとして残す．
一例として，以下の自動アノテーションを考える．



ここでは，6番目の文にある固有表現「大沢直衛」が，2番目の文にある固有表現「椎谷堀家」に所属すると示している．
文脈から，以下の情報が汲み取れる．

- 「大沢直衛」は「堀　直宥」の五男（文6）
- 「堀　直宥」は「椎谷堀家」に所属する（文1，文2）

さらに，「子は親と同じ家系に所属することがよくある」という人間の一般常識がある．
よって，文脈での繋がり及び常識により，（大沢直衛，所属団体，椎谷堀家）が推測できることから，この関係を正しいとして残す．

(2) 根拠文の提示

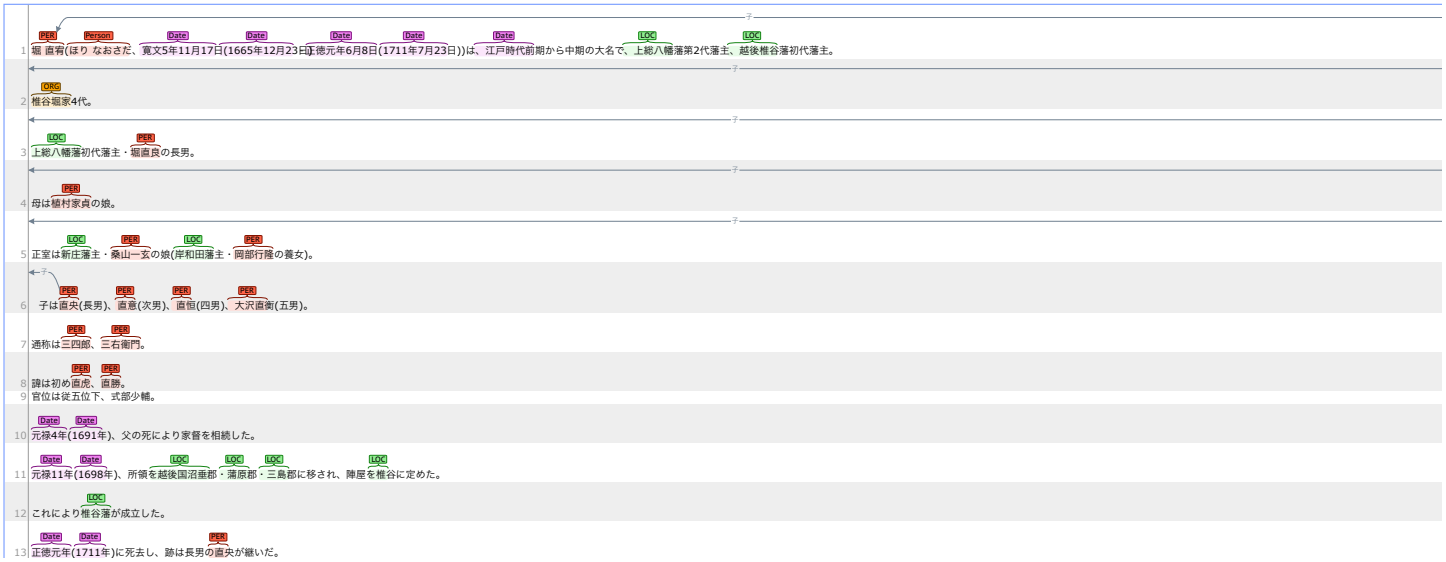
正しい関係（head, rel, tail）に対して，それを裏付ける根拠を文単位で提示されたい．
根拠文の提示は，該当関係のラベルをダブルクリックし，Notes欄に根拠文のIDをコンマ区切りで記入する形で行う．
ただし，固有表現 `head` と `tail` が所在する文を根拠文と黙認する．このため，`head` と `tail` が所在する文だけで `rel` を推測できる場合は，根拠文の提示が不要とする．

根拠文の提示が不要な場合

例えば，以下の関係アノテーションを考える．

ここでは，関係（堀 直宥，生年月日，寛文5年11月17日）が示されている．常識上，「氏名（生年月日－没年月日）」という一般的な表記形式があることから，この関係を正しいとして残す．
我々人間は文1で当関係を推測できたことから，当関係の根拠文は文1だけである．一方，`head`である「堀 直宥」と`tail`である「寛文5年11月17日」両方が所在する文1は，既に根拠文として黙認されるため，根拠文の提示は不要である．

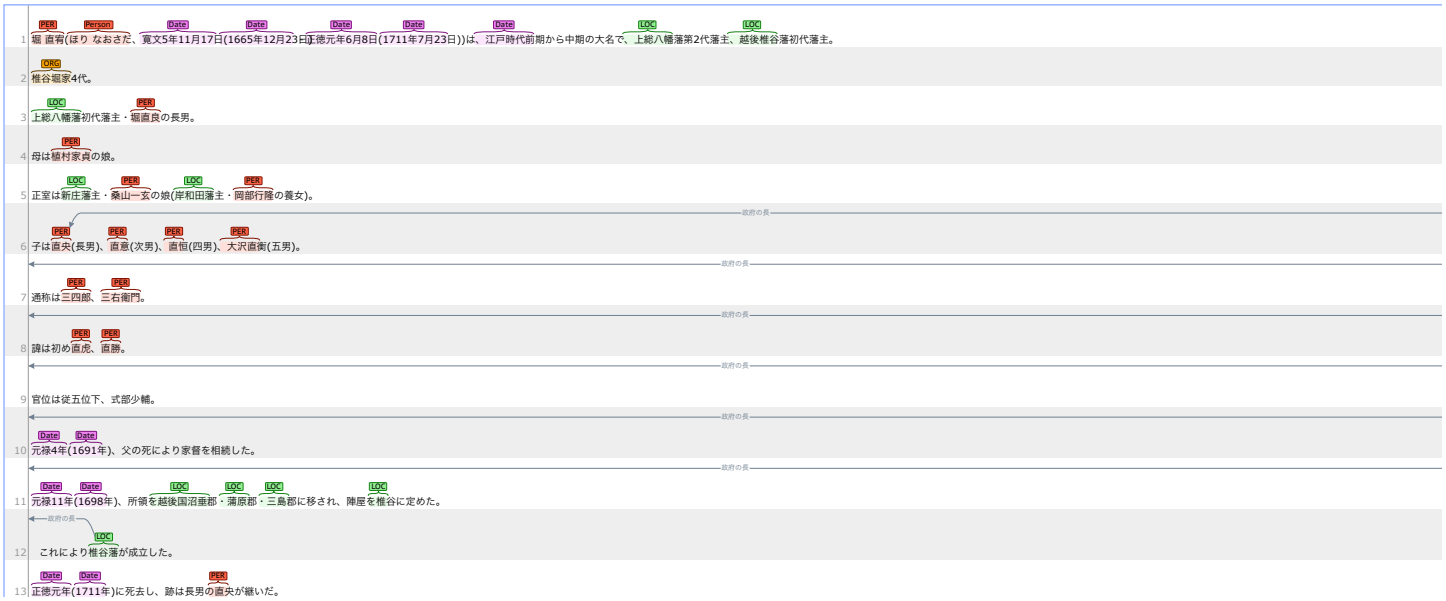
さらに以下の関係アノテーションを考える．



ここでは，文6にある固有表現「直央」は文1にある固有表現「堀 直宥」の子である関係が示されている．この関係は，両固有表現がそれぞれ所在する文6と文1から直接読み取れるため，根拠文の提示は不要である．

根拠文の提示が必要な場合

以下のアノテーションを考える．



副節：残すべき関係で説明されたように，ここの関係（椎谷藩，政府の長，直央）は，文1・6・13から推測できる．このため，`head`である「椎谷藩」が所在する文12と`tail`である「直央」が所在する文6以外にも，文1と文13を根拠文として提示する必要がある．

具体的には，関係ラベル「政府の長」をダブルクリックし，根拠文を以下のようにNotes欄に記入する．

(3) 新規関係の追加

新規関係の追加作業は，全ての関係を見つけ出すことを目標としない．以下いずれの条件を満たせば，作業完了とする．

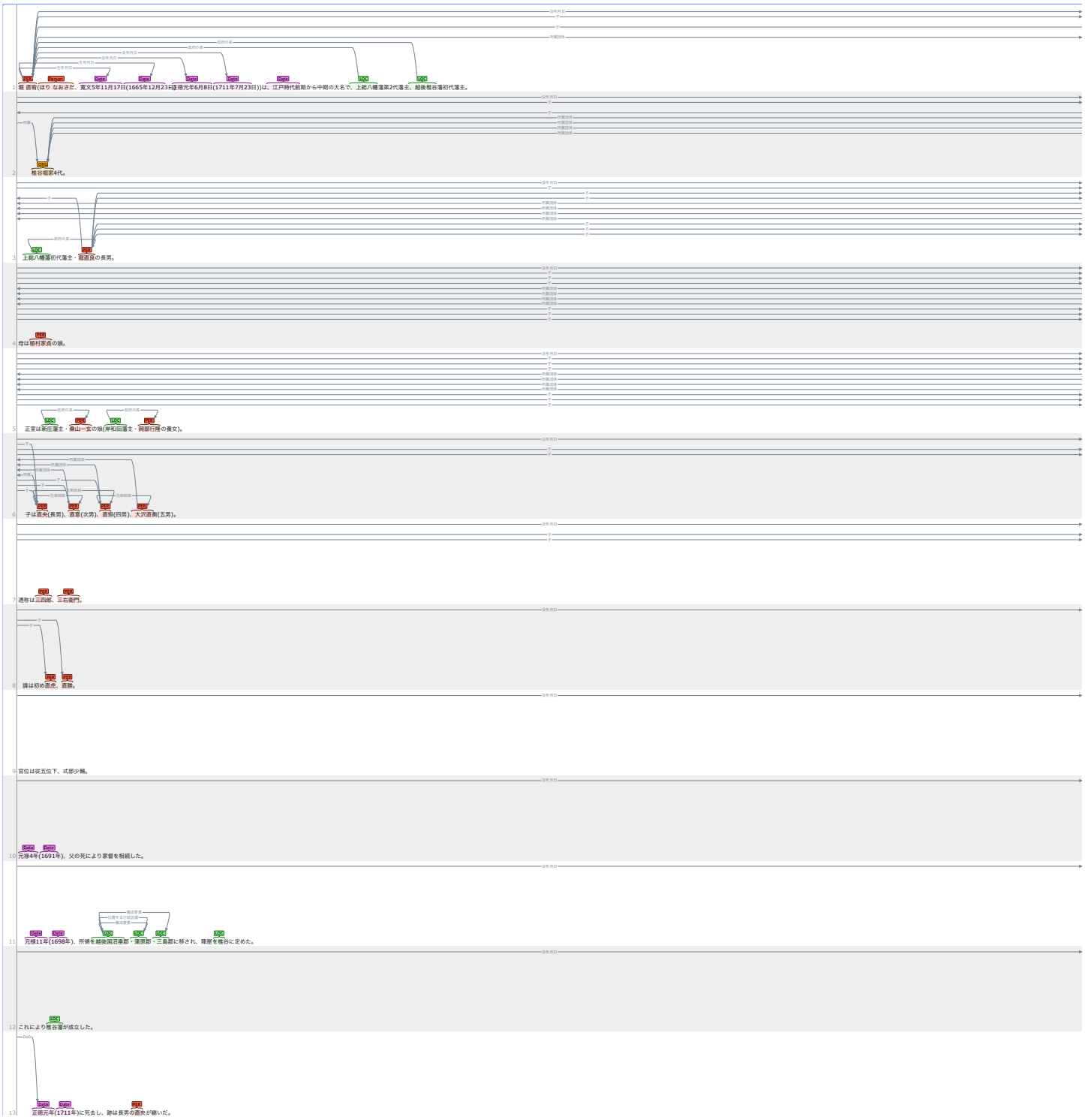
+ タイムリミットを超えた
+ 関係ラベルを一定数まで追加した
+ 全ての関係ラベルが付与済みであると判断した

自動認識されなかった関係を全ての関係アノテーションをまとめたファイルに追加する（まとめファイルは作業者の都合の良いように作成して頂きたい）．
なお，共参照関係のある固有表現は，共参照セットの内任意の固有表現と紐付けていれば良い．

例えば，文書0002の関係まとめファイル `0002_all` の中身は以下とする．

このファイルでは，以下の関係が欠落していると判断し，関係及び根拠文のアノテーションを追加する（タイムリミット：3分）．

+ （堀　直宥，所属団体，椎谷堀家）
+ （直央，所属団体，椎谷堀家）
+ （直意，所属団体，椎谷堀家）
+ （直恒，所属団体，椎谷堀家）
+ （大沢直衡，所属団体，椎谷堀家）
+ （大沢直衡，兄弟姉妹，直恒）
+ （上総八幡藩，政府の長，堀　直宥）
+ （越後椎谷藩，政府の長，堀　直宥）
+ （上総八幡藩，政府の長，堀直良）
+ （新庄藩，政府の長，桑山一玄）
+ （岸和田藩，政府の長，岡部行隆）



5. 関係ラベルセット

関係ラベルセットの設計は，自由利用可能・多言語などを特徴とする構造化データのデータベース・ウィキデータに基づく．
ウィキデータでは項目（実体）間の関係を表すものとして，「国籍」・「所属団体」・「所在地」などのプロパティが定義されている（詳細はWikidata:はじめにを参照）．

作業対象となる関係ラベルは，ウィキデータのプロパティから以下**28種類**を選出した．抽出された関係情報を三つ組の形（head, rel, tail）とし，固有表現`head`と`tail`の種類によって付与できる関係`rel`の種類が制限される．

ORG＝組織名（ORGANIZATION），PER＝人名（PERSON），LOC＝地名（LOCATION），ART＝固有物名（ARTIFACT），DATE＝日付表現（DATE）．各行のWikidata IDをクリックすると，対応するWikidataページに遷移できる．さらに関係の説明や具体例を確認したい場合は遷移先に参照されたい．

	Relation ID (RID)	label	Wikidata ID	head	tail	description	example
位置関係（Physical）	01	政庁所在地（Capital）	P1376	LOC	LOC	tailはheadを行政の中心とする国・州・行政区画.	（東京，政庁所在地，日本）
	02	位置する行政区画（AdministrativeLocation）	P131	ORG/LOC/ART	LOC	tailはheadが位置している行政区画.	（東京ディズニーランド，位置する行政区画，千葉県）
	03	所在地（Location）	P276	ORG/LOC/ART	LOC	移動可能な物がある場所，構造物の所在地，出来事の発生地．tailは行政区画でないもの.	（鳥獣人物戯画，所在地，高山寺）
	04	活動地（WorkLocation）	P937	ORG/PER	LOC	headは人物・組織（芸術家など）で，tailはそれが活動（作品の創作など）	（フィリッポ・ブルネレスキ，活動地，フィレンツェ）

	Relation ID (RID)	label	Wikidata ID	head	tail	description	example
						をした場所.	
固有属性（General Affiliation）	05	国籍（CountryOfCitizenship）	P27	PER	LOC	tailはheadを自国の市民として認めている国.	（ナポレオン・ボナパルト，国籍，フランス第一帝政）
	06	生年月日（DateOfBirth）	P569	PER	DATE	tailはheadの人物が生まれた日付. tailが完全な日付（XXXX年XX月XX日）である必要はない.	（アイザック・ニュートン，生年月日，1642年）
	07	没年月日（DateOfDeath）	P570	PER	DATE	tailはheadの人物が死亡した日付. tailが完全な日付である必要はない.	（アイザック・ニュートン，没年月日，1727年3月20日）
	08	生地（PlaceOfBirth）	P19	PER	LOC	tailはheadの人物が生まれた場所.	（坂本龍馬，生地，土佐国土佐郡）
	09	没地（PlaceOfDeath）	P20	PER	LOC	tailはheadの人物が死亡した場所.	（坂本龍馬，没地，京都）
	10	前項（Follows）	P155	ORG/LOC/ART	ORG/LOC/ART	headが国家・組織の場合，tailがその前身； headが創作物の場合，tailがその前巻・前作.	（トルコ，前項，オスマン帝国）
人間ー社会関係（Personal-Social）	11	子（Child）	P40	PER	PER	headはtailの子. （wikidataでの定義と方向が異なる）	（イエス・キリスト，子，聖母マリア）
	12	兄弟姉妹（Sibling）	P3373	PER	PER	tailはheadの兄弟姉妹. （headとtailは対称で方向なし）	（源義経，兄弟姉妹，源頼朝）
	13	配偶者（Spouse）	P26	PER	PER	tailはheadの配偶者. （headとtailは対称で方向なし）	（ジョー・バイデン，配偶者，ジル・バイデン）
	14	参加イベント（PartifipantIn）	P1344	ORG/PER/LOC	ART	tailはhead（人物や組織・グループ）が参加・出場したイベント.	（大谷翔平，参加イベント，ワールド・ベースボール・クラシック）
全体ー部分関係（Part-Whole）	15	所属団体（MemberOf）	P463	ORG/PER/LOC	ORG	headは人物・組織で， tailはそれが所属している団体.	（二宮和也，所属団体，嵐）
	16	構成要素（PartOf）	P361	ORG/LOC/ART	ORG/PER/LOC/ART	headはtailの構成要素. R15に属さない全体ー部分関係に付与する.	（副都心線，構成要素，東京メトロ）
組織所属関係（Organization Affiliation）	17	政府の長（HeadOfGovernment）	P6	ORG/LOC	PER	headは行政府（国・都道府県・市・町その他の自治体）か政府機関で， tailはその長	（ドイツ，政府の長，オラフ・ショルツ）
	18	所有者（OwnedBy）	P127	LOC/ART	ORG/PER/LOC	headは人工物で，tailはその所有者（人・組織・自治体・国家）.	（エッフェル塔，所有者，パリ）
	19	創設者（FoundedBy）	P112	ORG/LOC	PER	headは組織・場所で， tailはその創設者または共同創設者.	（第一国立銀行，創設者，渋沢栄一）
	20	雇い主（Employer）	P108	PER	ORG/PER	headは人物で，tailはそれを雇用する人・組織.	（ニール・アームストロング，雇い主，NASA）
	21	運営元（Operator）	P137	LOC/ART	ORG/PER/LOC	headはサービス・施設・設備で， tailはそれを運営する人・組織・自治体・国家.	（山手線，運営元，JR東日本）
	22	出身校（EducatedAt）	P69	PER	ORG	tailはheadの出身校.	（白川英樹，出身校，東京工業大学）
知的財産が関わる関係（Intellect-Related）	23	受賞（AwardReceived）	P166	ORG/PER	ART	headは人物・組織で，tailはそれが受けた賞・表彰.	（川端康成，受賞，ノーベル文学賞）
	24	作者（Creator）	P170	ORG/PER/LOC/ART	ORG/PER	headは作品あるいは（架空の）人物・組織・場所・固有物で，tailはその作者.	（舞姫，作者，森鴎外）
	25	演者（Performer）	P175	PER/ART	ORG/PER	headは作中人物或いは音楽作品で， tailはそれを演じた（演奏した）人・組織.	（徳川家康，演者，松本潤）
	26	発行元（Publisher）	P123	ART	ORG/PER	headは本・出版物・ゲーム・ソフトウェアで，tailはその発行者・出版社・発売元・販売元	（ポケモンSV，発行元，株式会社ポケモン）
	27	登場する作品（PresentInWork）	P1441	ORG/PER/LOC/ART	ART	headは（架空の）人物・組織・場所・固有物で，tailはそれが登場する作品.	（諸葛孔明，登場する作品，三国志）
	28	対応機種（Platform）	P400	ART	ART	headはソフトウェアで， tailはそれの開発及びリリースされた機種	（ゼルダの伝説ティアーズオブザキングダム，対応機種，Nintendo Switch）

6. 専門知識と常識

今回の作業では「常識に準しているかどうか」を判断基準の一つとするが，「専門知識」か「常識」かの判断は作業者個々の直感に委ねる．

以上

Bibliography

[1] Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. Can knowledge graphs reduce hallucinations in LLMs? : A survey. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3947–3960, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.naacl-long.219.

[2] Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. Generating high quality proposition Banks for multilingual semantic role labeling. In Chengqing Zong and Michael Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1039. URL https://aclanthology.org/P15-1039.

[3] Peggy M. Andersen, Philip J. Hayes, Steven P. Weinstein, Alison K. Huettner, Linda M. Schmandt, and Irene B. Nirenburg. Automatic extraction of facts from press releases to generate news stories. In *Third Conference on Applied Natural Language Processing*, pages 170–177, Trento, Italy, March 1992. Association for Computational Linguistics. doi: 10.3115/974499.974531. URL https://aclanthology.org/A92-1024.

[4] Ron Artstein. *Inter-annotator Agreement*, pages 297–313. Springer Netherlands, Dordrecht, 2017. ISBN 978-94-024-0881-2. doi:

10.1007/978-94-024-0881-2_11. URL https://doi.org/10.1007/978-94-024-0881-2_11.

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

[6] Jinheon Baek, Alham Aji, and Amir Saffari. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In Estevam Hruschka, Tom Mitchell, Sajjadur Rahman, Dunja Mladenić, and Marko Grobelnik, editors, *Proceedings of the First Workshop on Matching From Unstructured and Structured Data (MATCHING 2023)*, pages 70–98, Toronto, ON, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.matching-1.7. URL https://aclanthology.org/2023.matching-1.7.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. URL https://api.semanticscholar.org/CorpusID:11212020.

[8] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://aclanthology.org/D19-1371.

[9] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.

[10] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null): 1137–1155, March 2003. ISSN 1532-4435.

[11] Victoria Bobicev and Marina Sokolova. Inter-annotator agreement in sentiment analysis: Machine learning perspective. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International*

Conference Recent Advances in Natural Language Processing, RANLP 2017, pages 97–102, Varna, Bulgaria, September 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_015. URL https://doi.org/10.26615/978-954-452-049-6_015.

[12] Richard W. Brislin. Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, 1(3):185–216, 1970. doi: 10.1177/135910457000100301. URL https://doi.org/10.1177/135910457000100301.

[13] 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 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[14] Haotian Chen, Bingsheng Chen, and Xiangdong Zhou. Did the models understand documents? benchmarking models for language understanding in document-level relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6418–6435, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.354. URL https://aclanthology.org/2023.acl-long.354.

[15] Tongfei Chen, Yunmo Chen, and Benjamin Van Durme. Hierarchical entity typing via multi-level learning to rank. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8465–8475, Online, July 2020. Association for Computational Linguis-

tics. doi: 10.18653/v1/2020.acl-main.749. URL https://aclanthology.org/2020.acl-main.749.

[16] Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. Frustratingly easy label projection for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.357. URL https://aclanthology.org/2023.findings-acl.357.

[17] Fei Cheng, Shuntaro Yada, Ribeka Tanaka, Eiji Aramaki, and Sadao Kurohashi. JaMIE: A pipeline Japanese medical information extraction system with novel relation annotation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3724–3731, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.397.

[18] 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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.249. URL https://aclanthology.org/2021.findings-acl.249.

[19] Barry Chiswick and Paul Miller. Linguistic distance: A quantitative measure of the distance between english and other languages. IZA Discussion Papers 1246, Institute of Labor Economics (IZA), 2004. URL https://EconPapers.repec.org/RePEc:iza:izadps:dp1246.

[20] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020. URL https://openreview.net/pdf?id=r1xMH1BtvB.

[21] Paolo Coletti and Marco Costantino. *Information Extraction in Finance*. Wit Press, 01 2008. ISBN 978-1-84564-146-7.

[22] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 160–167, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390177. URL https://doi.org/10.1145/1390156.1390177.

[23] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[24] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

[25] Wanyun Cui, Yanghua Xiao, Haixun Wang, Yangqiu Song, Seung-won Hwang, and Wei Wang. KBQA: learning question answering over QA corpora and knowledge bases. *CoRR*, abs/1903.02419, 2019. URL http://arxiv.org/abs/1903.02419.

[26] Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290, 2022. doi: 10.1162/tacl_a_00460. URL https://aclanthology.org/2022.tacl-1.16.

[27] Julien Delaunay, Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Georgeta Bordea, Nicolas Sidere, and Antoine Doucet. A comprehensive survey of document-level relation extraction (2016-2023), 2023.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language un-

derstanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[29] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The automatic content extraction (ACE) program – tasks, data, and evaluation. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf.

[30] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A survey on in-context learning, 2023.

[31] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1045. URL https://aclanthology.org/D18-1045.

[32] Manaal Faruqui and Shankar Kumar. Multilingual open relation extraction using cross-lingual projection. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1356, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1151. URL https://aclanthology.org/N15-1151.

[33] Jenny Rose Finkel and Christopher D. Manning. Nested named entity recognition. In Philipp Koehn and Rada Mihalcea, editors, *Proceedings of*

the 2009 Conference on Empirical Methods in Natural Language Processing, pages 141–150, Singapore, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/D09-1015.

[34] Robert Gaizauskas and Yorick Wilks. Information extraction: Beyond document retrieval. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 3, Number 2, August 1998*, pages 17–60, August 1998. URL https://aclanthology.org/O98-4002.

[35] Ralph Grishman. Twenty-five years of information extraction. *Natural Language Engineering*, 25(6):677–692, 2019. doi: 10.1017/S1351324919000512.

[36] Ralph Grishman and Beth Sundheim. Message Understanding Conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996. URL https://aclanthology.org/C96-1079.

[37] Jia Guo, Stanley Kok, and Lidong Bing. Towards integration of discriminability and robustness for document-level relation extraction. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2606–2617, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.191. URL https://aclanthology.org/2023.eacl-main.191.

[38] Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 221–231, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1021. URL https://aclanthology.org/P17-1021.

[39] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Los Alamitos, CA, USA, jun 2016.

IEEE Computer Society. doi: 10.1109/CVPR.2016.90. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.90.

[40] Leonhard Hennig, Philippe Thomas, and Sebastian Möller. MultiTACRED: A multilingual version of the TAC relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.210. URL https://aclanthology.org/2023.acl-long.210.

[41] Kevin Huang, Peng Qi, Guangtao Wang, Tengyu Ma, and Jing Huang. Entity and evidence guided document-level relation extraction. In Anna Rogers, Iacer Calixto, Ivan Vulić, Naomi Saphra, Nora Kassner, Oana-Maria Camburu, Trapit Bansal, and Vered Shwartz, editors, *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 307–315, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.repl4nlp-1.30. URL https://aclanthology.org/2021.repl4nlp-1.30.

[42] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023.

[43] 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 Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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)*, pages 998–1004, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.126. URL https://aclanthology.org/2021.acl-short.126.

[44] Quzhe Huang, Shibo Hao, Yuan Ye, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. Does recommend-revise produce reliable annotations? an analysis on missing instances in DocRED. In Smaranda Muresan,

Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6241–6252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.432. URL https://aclanthology.org/2022.acl-long.432.

[45] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging, 2015.

[46] Filip Ilievski, Daniel Garijo, Hans Chalupsky, Naren Teja Divvala, Yixiang Yao, Craig Rogers, Ronpeng Li, Jun Liu, Amandeep Singh, Daniel Schwabe, and Pedro Szekely. KGTK: A toolkit for large knowledge graph manipulation and analysis. In *International Semantic Web Conference*, pages 278–293. Springer, 2020. URL https://arxiv.org/pdf/2006.00088.pdf.

[47] Heng Ji and Ralph Grishman. Knowledge base population: Successful approaches and challenges. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1148–1158, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1115.

[48] Robin Jia, Cliff Wong, and Hoifung Poon. Document-level n-ary relation extraction with multiscale representation learning. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3693–3704, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1370. URL https://aclanthology.org/N19-1370.

[49] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020. doi: 10.1162/tacl_a_00300. URL https://aclanthology.org/2020.tacl-1.5.

[50] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, USA, 1st edition, 2000. ISBN 0130950696.

[51] Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. Knowledge-consistent dialogue generation with knowledge graphs. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*, 2022. URL https://openreview.net/forum?id=McHtKDi5h9.

[52] Keshav Kolluru, Muqeeth Mohammed, Shubham Mittal, Soumen Chakrabarti, and Mausam . Alignment-augmented consistent translation for multilingual open information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2502–2517, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.179. URL https://aclanthology.org/2022.acl-long.179.

[53] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3230.

[54] Seth Kulick, Ann Bies, and Justin Mott. Inter-annotator agreement for ERE annotation. In Teruko Mitamura, Eduard Hovy, and Martha Palmer, editors, *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 21–25, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-2904. URL https://aclanthology.org/W14-2904.

[55] Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore, December 2023. Association for

Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.28. URL https://aclanthology.org/2023.emnlp-demo.28.

[56] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

[57] Mohamed Yassine Landolsi, Lobna Hlaoua, and Lotfi Ben Romdhane. Information extraction from electronic medical documents: state of the art and future research directions. *Knowl. Inf. Syst.*, 65(2):463–516, nov 2022. ISSN 0219-1377. doi: 10.1007/s10115-022-01779-1. URL https://doi.org/10.1007/s10115-022-01779-1.

[58] Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation, 2023.

[59] 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, 05 2016. ISSN 1758-0463. doi: 10.1093/database/baw068. URL https://doi.org/10.1093/database/baw068. baw068.

[60] Junpeng Li, Zixia Jia, and Zilong Zheng. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5495–5505, December 2023. doi: 10.18653/v1/2023.emnlp-main.334. URL https://aclanthology.org/2023.emnlp-main.334.

[61] Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=cPgh4gWZlz.

[62] Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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)*, pages 5834–5846, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.453. URL https://aclanthology.org/2021.acl-long.453.

[63] Xiao Liu, Heyan Huang, and Yue Zhang. Open domain event extraction using neural latent variable models. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2860–2871, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1276. URL https://aclanthology.org/P19-1276.

[64] 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. *ArXiv*, abs/1907.11692, 2019. URL https://arxiv.org/abs/1907.11692.

[65] Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1598. URL https://aclanthology.org/P19-1598.

[66] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[67] Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. A general framework for information extraction using dynamic span graphs. In Jill Burstein, Christy Doran, and Thamar Solorio,

editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3036–3046, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1308. URL https://aclanthology.org/N19-1308.

[68] Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning, 2024.

[69] Youmi Ma, Bhushan Kotnis, Carolin Lawrence, Goran Glavaš, and Naoaki Okazaki. Improving cross-lingual transfer for open information extraction with linguistic feature projection. In Duygu Ataman, editor, *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 125–138, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.mrl-1.11. URL https://aclanthology.org/2023.mrl-1.11.

[70] 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*, pages 1971–1983, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.145. URL https://aclanthology.org/2023.eacl-main.145.

[71] Mike Mintz, Steven Bills, Rion Snow, and Daniel 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 Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/P09-1113.

[72] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using LSTMs on sequences and tree structures. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1105. URL https://aclanthology.org/P16-1105.

[73] Jian Ni, Georgiana Dinu, and Radu Florian. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1135. URL https://aclanthology.org/P17-1135.

[74] OpenAI. Gpt-4 technical report, 2024.

[75] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[76] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, page 1–20, 2024. ISSN 2326-3865. doi: 10.1109/tkde.2024.3352100. URL http://dx.doi.org/10.1109/TKDE.2024.3352100.

[77] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. Cross-sentence n-ary relation extraction with graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115, 2017. doi: 10.1162/tacl_a_00049. URL https://aclanthology.org/Q17-1008.

[78] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL https://aclanthology.org/P19-1493.

[79] Chris Quirk and Hoifung Poon. Distant supervision for relation extraction beyond the sentence boundary. In Mirella Lapata, Phil Blunsom, and

Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1171–1182, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1110.

[80] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. In *OpenAI Technical Report*, 2018. URL https://api.semanticscholar.org/CorpusID:49313245.

[81] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. In *OpenAI Technical Report*, 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

[82] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *ECML/PKDD*, 2010. URL https://api.semanticscholar.org/CorpusID:2386383.

[83] Dan Roth and Wen-tau Yih. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8, Boston, Massachusetts, USA, May 6 - May 7 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-2401.

[84] Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In Suresh Manandhar and Deniz Yuret, editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/S13-2056.

[85] Satoshi Sekine and Hitoshi Isahara. IREX: IR & IE evaluation project in Japanese. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, May

2000. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2000/pdf/27.pdf.

[86] Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-0812. URL https://aclanthology.org/W15-0812.

[87] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April 2012. Association for Computational Linguistics. URL https://aclanthology.org/E12-2021.

[88] Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. Document-level relation extraction with adaptive focal loss and knowledge distillation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.132. URL https://aclanthology.org/2022.findings-acl.132.

[89] Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. Revisiting DocRED - addressing the false negative problem in relation extraction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.580. URL https://aclanthology.org/2022.emnlp-main.580.

[90] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In

Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pages 142–147, 2003. URL https://aclanthology.org/W03-0419.

[91] Bayu Distiawan Trisedya, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 229–240, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1023. URL https://aclanthology.org/P19-1023.

[92] Nobuhiro Ueda, Kazumasa Omura, Takashi Kodama, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. KWJA: A unified Japanese analyzer based on foundation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 538–548, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-demo.52. URL https://aclanthology.org/2023.acl-demo.52.

[93] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[94] Patrick Verga, Emma Strubell, and Andrew McCallum. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 872–884, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1080. URL https://aclanthology.org/N18-1080.

[95] Jesse Vig, Alexander Fabbri, Wojciech Kryscinski, Chien-Sheng Wu,

and Wenhao Liu. Exploring neural models for query-focused summarization. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1455–1468, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.109. URL https://aclanthology.org/2022.findings-naacl.109.

[96] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, sep 2014. ISSN 0001-0782. doi: 10.1145/2629489. URL https://doi.org/10.1145/2629489.

[97] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1585. URL https://aclanthology.org/D19-1585.

[98] 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)*, pages 15566–15589, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.868. URL https://aclanthology.org/2023.acl-long.868.

[99] Yilin Wen, Zifeng Wang, and Jimeng Sun. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

[100] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama

Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

[101] Ye Wu, Ruibang Luo, Henry C. M. Leung, Hing-Fung Ting, and Tak-Wah Lam. Renet: A deep learning approach for extracting gene-disease associations from literature. In Lenore J. Cowen, editor, *Research in Computational Molecular Biology*, pages 272–284, Cham, 2019. Springer International Publishing. URL https://link.springer.com/chapter/10.1007/978-3-030-17083-7_17.

[102] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016.

[103] Chaojun Xiao, Yuan Yao, Ruobing Xie, Xu Han, Zhiyuan Liu, Maosong Sun, Fen Lin, and Leyu Lin. Denoising relation extraction from document-level distant supervision. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3683–3688, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.300. URL https://aclanthology.org/2020.emnlp-main.300.

[104] Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, and Jiawei Han. SAIS: Supervising and augmenting intermediate steps for document-level relation extraction. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2395–2409, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.171. URL https://aclanthology.org/2022.naacl-main.171.

[105] 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 Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 257–268, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.23. URL https://aclanthology.org/2022.findings-acl.23.

[106] Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14149–14157, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17665.

[107] Wang Xu, Kehai Chen, and Tiejun Zhao. Document-level relation extraction with reconstruction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14167–14175, May 2021. doi: 10.1609/aaai.v35i16.17667. URL https://ojs.aaai.org/index.php/AAAI/article/view/17667.

[108] Wang Xu, Kehai Chen, Lili Mou, and Tiejun Zhao. Document-level relation extraction with sentences importance estimation and focusing. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2920–2929, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.212. URL https://aclanthology.org/2022.naacl-main.212.

[109] Bishan Yang and Tom M. Mitchell. Joint extraction of events and entities within a document context. In Kevin Knight, Ani Nenkova, and Owen

Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 289–299, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1033. URL https://aclanthology.org/N16-1033.

[110] Jie Yang and Yue Zhang. NCRF++: An open-source neural sequence labeling toolkit. In Fei Liu and Thamar Solorio, editors, *Proceedings of ACL 2018, System Demonstrations*, pages 74–79, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-4013. URL https://aclanthology.org/P18-4013.

[111] Soyoung Yang, Minseok Choi, Youngwoo Cho, and Jaegul Choo. HistRED: A historical document-level relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3207–3224, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.180. URL https://aclanthology.org/2023.acl-long.180.

[112] 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 Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1074. URL https://aclanthology.org/P19-1074.

[113] David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, 2001. URL https://aclanthology.org/H01-1035.

[114] Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. Improved neural relation detection for knowledge base question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 571–581, Vancouver, Canada, July 2017. Association for

Computational Linguistics. doi: 10.18653/v1/P17-1053. URL https://aclanthology.org/P17-1053.

[115] 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 *International Joint Conference on Artificial Intelligence*, 2020. URL https://api.semanticscholar.org/CorpusID:220484624.

[116] Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. Double graph based reasoning for document-level relation extraction. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.127. URL https://aclanthology.org/2020.emnlp-main.127.

[117] Shuang Zeng, Yuting Wu, and Baobao Chang. SIRE: Separate intra- and inter-sentential reasoning for document-level relation extraction. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 524–534, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.47. URL https://aclanthology.org/2021.findings-acl.47.

[118] Thomas Zenkel, Joern Wuebker, and John DeNero. End-to-end neural word alignment outperforms GIZA++. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1605–1617, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.146. URL https://aclanthology.org/2020.acl-main.146.

[119] Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. Document-level relation extraction as semantic segmentation. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3999–4006. International Joint Conferences on Artificial

Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/551. URL https://doi.org/10.24963/ijcai.2021/551. Main Track.

[120] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45, 2017. URL https://nlp.stanford.edu/pubs/zhang2017tacred.pdf.

[121] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Learning deep bilinear transformation for fine-grained image representation. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc. URL https://dl.acm.org/doi/10.5555/3454287.3454672.

[122] Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. Why does chatgpt fall short in providing truthful answers?, 2023.

[123] Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. ConNER: Consistency training for cross-lingual named entity recognition. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8438–8449, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.577. URL https://aclanthology.org/2022.emnlp-main.577.

[124] 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*, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17717/17524.

Publication List

Journals (Peer-Reviewed)

1. Youmi Ma, An Wang, and 岡崎 直観. 文書レベル関係抽出における根拠認識の統合. 自然言語処理, 31(1):to appear, March 2024.

2. Youmi Ma, Tatsuya Hiraoka, and Naoaki Okazaki.**Named Entity Recognition and Relation Extraction Using Enhanced Table Filling by Contextualized Representations**. 自然言語処理, 29(1):187–223, March 2022.

International Conferences (Peer-Reviewed)

1. Youmi Ma, An Wang, and Naoaki Okazaki. **Building a Japanese Document-Level Relation Extraction Dataset Assisted by Cross-Lingual Transfer.** In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), pages (to appear), Turin, Italy, May 2024.

2. Youmi Ma, Bhushan Kotnis, Carolin Lawrance, Goran Glavaš, and Naoaki Okazaki. **Improving Cross-Lingual Transfer for Open Information Extraction with Linguistic Feature Projection**. In The 3rd Multilingual Representation Learning Workshop (MRL), pages 125-138, Singapore, December 2023.

3. 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: Main Volume (EACL), pages 1971–1983, Dubrovnik, Croatia, May 2023.

4. 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), pages 11–21, Dublin, Ireland, May 2022.

5. An Wang, Junfeng Jiang, Youmi Ma, Ao Liu, and Naoaki Okazaki. **Generative Data Augmentation for Aspect Sentiment Quad Prediction**. In Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM), pages 128–140, Toronto, Canada, July 2023.

6. Hsuan-Yu Kuo, Youmi Ma, and Naoaki Okazaki. **Annotating Entity and Causal Relationships on Japanese Vehicle Recall Information**. In Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation (PACLIC), pages 783–791, Manila, Philippines, October 2022.

7. Zixia Jia, Youmi Ma, Jiong Cai, and Kewei Tu. **Semi-Supervised Semantic Dependency Parsing Using CRF Autoencoders**. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pages 6795–6805, Online, July 2020.

Domestic Conferences (Non Peer-Reviewed)

1. Youmi Ma, An Wang, 岡崎 直観. 日本語文書レベル関係抽出コーパスの構築. 第18回NLP若手の会シンポジウム, S5-P19, 2023年8月.

2. Youmi Ma, An Wang, 岡崎 直観. 文書レベル関係抽出における根拠認識の統合. 言語処理学会第29回年次大会 (NLP2023), B3-3, pp. 605–610, 2023年3月.

3. Youmi Ma, An Wang, 岡崎 直観. 文書レベル関係抽出における人間と注意機構の根拠文の対応付け. 第17回NLP若手の会シンポジウム, P2-03, 2022年8月.

4. Youmi Ma, 平岡 達也, 岡崎 直観. 畳み込みニューラルネットワークを用いた表ラベリングによる固有表現認識と関係抽出. 言語処理学会第28回年次大会 (NLP2022), pp. 1197–1202, 2022年3月.

5. Youmi Ma, 平岡 達也, 岡崎 直観. **BERT**を用いた**Table-Filling**による固有表現抽出と関係抽出. 言語処理学会第27回年次大会 (NLP2021), pp. 1274–1279, 2021年3月.

6. 服部 翔, Youmi Ma, 岡崎 直観. クエリ指向要約におけるクエリと要約の統合的な生成. 言語処理学会第29回年次大会 (NLP2023), H5-2, pp. 1244–1249, 2023年3月.

Domestic Articles

1. Youmi Ma. 「文書レベル関係抽出における根拠認識の統合」の完成まで. 自然言語処理, 30(3):1088–1093, 2023年9月.