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
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	情報理工 知能情報	系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of	（ 工学 ）
学生氏名： Student's Name	MA Youmi		審査員主査： Chief Examiner	岡崎直観	

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

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.

For model construction, the goal is to obtain a DocRE model with high performance. To this end, this study proposes a training strategy named Document-level Relation Extraction with Evidence-guided Attention Mechanism (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 Japanese Document-level Relation Extraction Dataset, 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 existing 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.