HUHKA at the NTCIR-15 QA Lab-PoliInfo-2 Entity Linking Task

Takuma Himori
Graduate School of Information Science and Technology, Hokkaido University, Japan
himori@ist.hokudai.ac.jp

Yasutomo Kimura
Otaru University of Commerce, Japan
RIKEN, Japan
kimura@res.otaru-uc.ac.jp

Kenji Araki
Faculty of Information Science and Technology, Hokkaido University, Japan
araki@ist.hokudai.ac.jp

ABSTRACT

The HUHKA team participated in the Entity Linking task of Question Answering Lab for Political Information 2 (QA Lab-PoliInfo-2) subtask of the NTCIR-15. This report describes the methods we implemented to solve the task and discusses the results. We extract mentions of “law name” with BERT and filters. Moreover, we link the mention to Wikipedia with knowledge bases such as Wikipedia and e-Gov.

TEAM NAME

HUHKA

SUBTASKS

Entity Linking Task

1 INTRODUCTION

The NTCIR-15’s Question Answering Lab for Political Information 2 [1](QA Lab-PoliInfo-2) deals with political information and establishes four tasks: stance classification, dialog summarization, entity linking, and topic detection.

Entity linking refers to the task of extracting mentions of specific entities in text and assigning unique identifiers to them from an underlying knowledge base such as Wikipedia.

In this report, we describe two of our entity linking methods: named entity recognition and named entity disambiguation. Moreover, the results are discussed.

2 RELATED WORK

Entity linking is generally divided into two types of tasks: mention extraction and disambiguation. Mentions are references to entities and expressions that we want to associate with entities.

In the mention extraction task, expressions associated with entities are extracted from the text. In general, named entity recognition techniques are used, and IOB2 tags[2] are often used to tag a range of mentions in the text.

Recently, a named entity recognition model using BERT[3] was proposed. DeepPavlov[4], an open-source library, provides a multilingual named entity recognition model that uses BERT.

In the disambiguation task, the first step is to generate candidates for the entities to be associated with the mentions. The generated candidates are ranked, and the entity is associated with the highest one with a mention.

For the disambiguation task, Yamada et al.[5] extend the skip-gram model[6][7] to two models, the link graph model and the anchor context model, to disambiguate the entities.

3 ENTITY LINKING TASK IN QA LAB-POLIINFO-2

The Entity Linking task in QA Lab-PoliInfo-2 consists of extracting a mention of “law name” from politicians’ statements, disambiguation, and linking the mention to a knowledge base. Given the local assembly member’s utterances, the systems extract a mention of “law name” and link the mention with the list of law names or Wikipedia.

Table 1 provides an example of test data for the Entity Linking task, whereas Table 2 provides an example of an answer sheet for the Entity Linking task.

As Table 1 shows, only the morphologically analyzed local assembly member’s utterances are provided as test data.

In the Entity Linking task, we extract the mentions, as described in Table 2, and indicate the part with the IOB2 tag.

Subsequently, the extracted mentions are linked to Wikipedia or a list of legal names.

This format is a TSV and is similar to the AIDA CoNLL-YAGO[8] dataset.

Table 1: Example of the Entity Linking test data

私			
の			
方			
から			
は			
IR			
法			
の			
導入			
に			
伴う			
変化			

Table 2: Example of the Entity Linking answer sheet

私			
の			
方			
から			
は			
IR	B	IR 法	特定複合観光施設区域の整備の推進に関する法律
法	I	IR 法	特定複合観光施設区域の整備の推進に関する法律
の			
導入			
に			
伴う			
変化			

4 OUR METHODS

We use a combination of named entity recognition and named entity disambiguation methods to solve the Entity Linking task. Our methods comprise two parts: named entity recognition and named entity disambiguation. In Section 4.1, we explain how to extract mentions of “law name“ with BERT and filters. Subsequently, Section 4.2 describes how to link the mention to Wikipedia with knowledge bases such as Wikipedia and e-Gov.

4.1 Named Entity Recognition

In this section, we explain our named entity recognition method. We extract mentions of “law name” with BERT, and filter the extracted mentions.

In Section 4.1.1, we explain the BERT model that is used. In Sections 4.1.2 and 4.1.3, we explain filter1 and filter2 which are used to filter the extracted mentions with BERT.

4.1.1 BERT. We used DeepPavlov, which is an open-source library, for the extraction of mentions. DeepPavlov[4] published a multilingual named entity recognition model using BERT, which was pretrained from the multilingual BERT using Ontonotes. We further fine tuned the model using the training data of QA Lab-PoliInfo-2 Entity Linking task datasets[1].

4.1.2 filter1. Filter1 procedure applies a filter to BERT’s input. Various mentions contain the word “法.” Therefore, if the sentence to be input to BERT does not contain the word “法,” it was filter with filter1 and set all outputs to “O.”

4.1.3 filter2. Filter2 procedure filters the mentions extracted by BERT with respect to a regular expression. The following regular expressions are used.

- 「.*[法|法律|法案|法制|法律案]$」

We extract the mentions that match this regular expression. If there is no match, the part is output as “O.”

4.2 Named Entity Disambiguation

In this section, we disambiguate the extracted mentions and link them to Wikipedia. We use the BERT described in Section 4 to perform the mention extraction. For the extracted mentions, we perform disambiguation using exact match, Wikipedia2Vec, mention-entity prior, and e-Gov.

4.2.1 exact match. Table 3 indicates that each Wikipedia article provides a unique page ID. A list of Wikipedia article titles paired with page IDs is included in the QA Lab-PoliInfo-2 Entity Linking Task dataset and can be used.

If an exact match occurs, that is, the extracted mentions and the Wikipedia title correspond to an exact match, the named entity disambiguation outputs the Wikipedia title. When no match occurs, it outputs “NIL.”

4.2.2 Wikipedia2Vec. Wikipedia2Vec[5] is a method proposed by Yamada et al. that can calculate the similarity between an input word and a Wikipedia article. We used Wikipedia2Vec to generate the output as the Wikipedia article title with the highest similarity to the extracted mentions. For the model of Wikipedia2Vec, we used the Japanese Wikipedia dump data from 2019.12.01 to train a 300-dimensional distributed representation.

Table 3: Example of Wikipedia Title and Page ID

Wikipedia Title	Page ID
独占禁止法	1343190
田山輝明	1343194
南極環境保護法	1343195
南極保護法	1343195
南極法	1343197

4.2.3 mention-entity prior. We select the top ranked entities based on the mention-entity prior $p(e|m)$ for a given entity e and a mention m. To compute this prior, we sum up hyperlink counts from Wikipedia to estimate probability $p(e|m)$. Ganea and Hofmann[9] generated candidate entities using Wikipedia and YAGO’s mention-entity prior. In the proposed method, we compute this mention entity prior, using the Japanese Wikipedia dump data from 2019.12.01.

4.2.4 e-Gov. E-Gov[1] provides a comprehensive search and guidance service for administrative information provided by ministries and agencies via the Internet. One of the systems provided by e-Gov is the law search system, which allows users to search and view Japanese laws on the website. The e-Gov law search system registers abbreviations of law names so that users can identify the formal and abbreviated names of laws. Table 4 presents an example of formal and abbreviated law names.

We use this formal name and abbreviation as a dictionary. As in section 4.2.1, if the extracted names are formal names that exist in the dictionary, we output the formal names. If the extracted mention is an abbreviation of a mention, the formal name of the abbreviation is output. If the mention does not exist in the dictionary, “NIL” is output.

5 RESULTS

Table 5 lists the obtained official formal run results, and Table 6 lists the obtained scores of the late submissions in formal run. NER stands for named entity recognition, whereas NED stands for named entity disambiguation.

The results are shown for the combination of named entity recognition and named entity disambiguation methods.

The results from the comparison between filter1 and filter2 showed that the use of filter2 improved all scores. This is probably because the wrong mention was extracted during the mention extraction process. Therefore, applying a filter to the extracted mentions is useful.

Subsequently, we compare Wikipedia2Vec and mention-entity prior. In the comparison, we found that the mention-entity prior was higher than Wikipedia2Vec, regardless of the named entity recognition method.

In addition, the combination of Wikipedia2Vec and mention-entity prior resulted in lower scores than when Wikipedia2Vec was implemented alone.

Moreover, the combination of Wikipedia2Vec and mention-entity prior obtained a score of 0.3980, which was similar to the score of

[1] https://www.e-gov.go.jp/

Table 4: Example of formal and abbreviated law names

Formal name	Abbreviation		
外務省設置法	中央省庁等改革関連法		
確定拠出年金法	日本版４０１ｋ法	ＤＣ法	
一般社団法人及び一般財団法人に関する法律	一般法人法	一般社団・財団法	一般社団・財団法人法
私的独占の禁止及び公正取引の確保に関する法律	独禁法	独占禁止法	
国債の発行等に関する省令	発行省令		
構造改革特別区域法	特区法	構造改革特区法	
広域臨海環境整備センター法	フェニックス法	フェニックス計画法	
小型自動車競走法	オートレース法		
国際連合平和維持活動等に対する協力に関する法律	ＰＫＯ協力法	国連平和協力法	ＰＫＯ法

0.3980 when Wikipedia2Vec was implemented alone. These results could be explained by Wikipedia2Vec having a larger effect on selecting entities, as we added the similarity of Wikipedia2Vec and the probability of the mention-entity prior as it is.

Finally, we compare the results of using e-Gov. Disambiguation using e-Gov alone produced lower results than mention-entity prior. However, when e-Gov was combined with other methods, all scores increased. Specifically, the combination of e-Gov and mention-entity prior showed the best results—a score of 0.6035. When combining e-Gov with other methods, e-Gov was used to disambiguate first. Then, the remaining mentions were disambiguated with other methods. By using dictionaries such as e-Gov to process mentions that could be reliably disambiguated, the results were considered better than those obtained by other methods when they were used alone.

Table 5: Scores in the formal run

NER Method	NED Method	Score
BERT + filter1	exact match	0.3247
BERT + filter1	mention-entity prior + Wikipedia2Vec	0.3980
BERT + filter1	e-Gov	0.4468
BERT + filter1	mention-entity prior	0.4747
BERT + filter1	e-Gov + mention-entity prior	0.4887
BERT + filter1 + filter2	e-Gov + mention-entity prior	**0.6035**

Table 6: Scores of the late submissions in the formal run

NER Method	NED Method	Score
BERT + filter1	Wikipedia2Vec	0.3980
BERT + filter1 + filter2	Wikipedia2Vec	0.5000
BERT + filter1 + filter2	e-Gov + Wikipedia2Vec	0.5130
BERT + filter1 + filter2	e-Gov	0.5518
BERT + filter1 + filter2	mention-entity prior	0.5863

6 CONCLUSIONS

We used a combination of named entity recognition and named entity disambiguation methods to solve the QA Lab-PoliInfo-2 Entity Linking Task. In the named entity recognition method, we found that filter2 was effective in filtering the extracted mentions with a regular expression.

For the named entity disambiguation method, all of the results showed higher scores when comparing e-Gov to other methods than when e-Gov was not used.

The best result was obtained using BERT, filter1, and filter2 for named entity recognition method and e-Gov and mention-entity prior for named entity disambiguation, which showed a score of 0.6035.

REFERENCES

[1] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Madoka Ishioroshi, Teruko Mitamura, Masaharu Yoshioka, Tomoyoshi Akiba, Yasuhiro Ogawa, Minoru Sasaki, Kenichi Yokote, Tatsunori Mori, Kenji Araki, Satoshi Sekine, and Noriko Kando. Overview of the ntcir-15 qa lab-poliinfo-2 task. *Proceedings of The 15th NTCIR Conference*, 12 2020.

[2] Erik F. Tjong, Kim Sang, and Jorn Veenstra. Representing text chunks. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway, June 1999. Association for Computational Linguistics.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] V. Mozharova and N. Loukachevitch. Two-stage approach in russian named entity recognition. In *2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*, pp. 1–6, Aug 2016.

[5] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. *arXiv preprint 1812.06280v3*, 2020.

[6] Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 3111–3119, USA, 2013. Curran Associates Inc.

[8] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 782–792, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[9] Octavian-Eugen Ganea and Thomas Hofmann. Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2619–2629, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.