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

Takuma Himori[1]

Yasutomo Kimura[2,3]

Kenji Araki[4]

[1] Graduate School of Information Science and Technology, Hokkaido University, Japan
[2] Otaru University of Commerce, Japan
[3] RIKEN, Japan
[4] Faculty of Information Science and Technology, Hokkaido University, Japan

Outline

1. Scores of each team
2. Our methods
    1. Named entity recognition（NER）methods
    2. Named entity disambiguation（NED）methods
3. Our results
4. Discussion
5. Conclusion

1. Scores of each team in the formal run

team	methods	Score
HUHKA	NER : BERT + filter（filter 1 and filter 2） NED : mention-entity prior+ e-Gov	0.6035
Forst	NER : rule-based + RNN NED : rule-based	0.3901
selt	NER : BERT NED : Wikipedia2Vec	0.2980
nukl	NER : Dictionary-based （NED : Dictionary-based）	0.2375

The score is the best results of each team.

We achieved the best result out of all the teams.

2. Our methods

We use a combination of named entity recognition and named entity disambiguation methods to solve the Entity Linking task.

Entity Linking task

Named entity recognition

We extract mentions of "law names" from local assembly member's utterances.

↓

Named entity disambiguation

We link the extracted mentions to Wikipedia title with knowledge bases i.e. Wikipedia and e-Gov[1].

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

2.1. Named entity recognition methods

We extract mentions of “law name” with BERT, and filter the extracted mentions using filter 1 and filter 2.

BERT

We use BERT model, which is available at DeepPavlov[1].
The model is a multilingual named entity recognition model, which was pretrained from the multilingual BERT using Ontonotes.
We further fine tuned the model on the training data of QA Lab-PoliInfo-2 Entity Linking task datasets.

Filter 1

If the sentence input into BERT does not contain the word “法”, it is filtered with filter 1 and all outputs are set to “O”.

Filter 2

We extract the mentions that match following regular expressions. If the mention does not match the following phrases, the output is “O”.

「.*[法|法律|法案|法制|法律案]¥$」

[1] 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.

2.2 Named entity disambiguation methods

We disambiguate the extracted mentions and link them to Wikipedia using exact match, Wikipedia2Vec, mention-entity prior, and e-Gov.

exact match

If the extracted mentions and the Wikipedia title corresponds to an exact match, the named entity disambiguation outputs the Wikipedia title.

Wikipedia2Vec[2]

We use Wikipedia2Vec to generate the output as the Wikipedia article title with the highest similarity to the extracted mentions.

mention-entity prior[3]

We select the top ranked entities based on the mention-entity prior $p(e|m)$, where e *is* a given entity and m *is* a mention.

e-Gov

We use the law search system provided by e-Gov. The system registers abbreviations of formal law names. We use these pairs of formal names and abbreviations as dictionary.

[2] 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

[3] Octavian-Eugen Ganea and Thomas Hofmann. Deep joint entity disambiguationwith 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.

3. Our results

NER methods	NED methods	Score
BERT + filter 1 + filter 2	mention-entity prior + e-Gov	0.6035
BERT + filter 1 + filter 2	mention-entity prior	0.5863
BERT + filter 1 + filter 2	e-Gov	0.5518
BERT + filter 1 + filter 2	Wikipedia2Vec + e-Gov	0.5130
BERT + filter 1 + filter 2	Wikipedia2Vec	0.5000
BERT + filter 1	mention-entity prior + e-Gov	0.4887
BERT + filter 1	mention-entity prior	0.4747
BERT + filter 1	e-Gov	0.4468
BERT + filter 1	Wikipedia2Vec	0.3980
BERT + filter 1	mention-entity prior + Wikipedia2Vec	0.3980
BERT + filter 1	exact match	0.3247

Scores in the formal run

Scores in the formal run（late submissions）

4. Discussion

・The combination methods of both the filter 1 and the filter 2 outperformed the results using only filter 1. This is probably because the wrong mention, like phrases which do not contain “法”, was extracted during the mention extraction process. These results showed filter 2 is also useful to remove noise.

・Disambiguation using e-Gov alone produced lower results than using mention-entity prior. However, when e-Gov was combined with other disambiguation methods, their scores increased.

・Specifically, the combination of e-Gov and mention-entity prior showed the best results—a score of 0.6035.

→ Using dictionaries such as e-Gov to process mentions that could be reliably disambiguated, the results of the combination methods were better than those obtained by other methods when they were used alone.

5. Conclusion

・We achieved the best score out of all the team.

・The combination of e-Gov and mention-entity prior showed the best results—a score of 0.6035.

・ Using Filter 2 and using e-Gov are useful to improve the score in this task.