Japanese Pointer Network based Entity Linker for Wikidata

Yuki Sawamura
Aoyama Gakuin University
Kanagawa, Japan
National Institute of Advanced
Industrial Science and Technology
Tokyo, Japan

Takeshi Morita
Aoyama Gakuin University
Kanagawa, Japan
National Institute of Advanced
Industrial Science and Technology
Tokyo, Japan
morita@it.aoyama.ac.jp

Shusaku Egami
National Institute of Advanced
Industrial Science and Technology
Tokyo, Japan

Takanori Ugai
Fujitsu Limited
Kanagawa, Japan
National Institute of Advanced
Industrial Science and Technology
Tokyo, Japan

Ken Fukuda
National Institute of Advanced
Industrial Science and Technology
Tokyo, Japan

ABSTRACT

Entity linking (EL) has attracted attention as a fundamental technology applicable to question answering and various other applications. State-of-the-art EL studies have focused on English with limited research on languages other than English. Current EL models are constructed using language models and knowledge graph embeddings, necessitating language-specific support for language models and knowledge graph embeddings. This study proposes Japanese PNEL (Pointer Network-based Entity Linker) by adapting the language-dependent embeddings in English PNEL. To achieve this, we analyzed the challenges in model construction, comparing Japanese and English PNEL from the perspective of embedding methods. Additionally, we translated the English dataset WebQSP into Japanese and evaluated our model using this dataset. The outcome of our study revealed that our Japanese PNEL outperformed state-of-the-art multilingual EL models when applied to WebQSP in Japanese EL.

CCS CONCEPTS

• **Computing methodologies → Information extraction**.

KEYWORDS

Knowledge Graph, Wikidata, Entity Linking

ACM Reference Format:
Yuki Sawamura, Takeshi Morita, Shusaku Egami, Takanori Ugai, and Ken Fukuda. 2023. Japanese Pointer Network based Entity Linker for Wikidata. In *Proceedings of (IJCKG '23)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
IJCKG '23, December 08–09,2023, Tokyo, Japan
© 2023 Association for Computing Machinery.
https://doi.org/XXXXXXX

1 INTRODUCTION

Entity Linking (EL) maps entity names in natural language sentences to resources in large knowledge graphs. Recently, EL systems have been gathering attention as a fundamental technology for question answering and other tasks.

Most EL studies focus on English, and few EL studies focus on languages other than English, including Japanese. Multilingual EL can perform for many languages but does not consider the unique characteristics of each language. In a comparative evaluation of the multilingual EL model and the English EL models for AIDA [10], the English EL model outperformed the multilingual EL model by 0.092 points in F1 values [12]. The multilingual EL model mGENRE [7] can run the Japanese EL but does not consider the Japanese characteristics. Therefore, Murawaki et al. [11] proposed a Japanese-specific EL task for Wikipedia, as the processing of Japanese is complicated by the fact that Japanese has a continuous character set. Recent EL studies [7, 2, 1, 19] use language models and knowledge graph embeddings to construct EL models and build EL models specialized in each language that needs specific language support.

We propose the Japanese PNEL (Pointer Network based Entity Linker) [2] as an EL model for Japanese. We selected the Pointer Network for constructing the Japanese EL model to enable analysis even when the required vector size varies according to language. First, we constructed the Japanese PNEL from the English PNEL model by changing language-dependent embeddings. Next, we translated the English datasets WebQSP [15], SimpleQuestions [4] and LC-QuAD2 [8] into Japanese. We carried out a comparative evaluation of the multilingual EL models and the Japanese PNEL on the translated WebQSP. As a result, the Japanese PNEL outperformed the multilingual EL model [7] regarding F1 values. In the ablation study, we analyzed language differences in embedding for English and Japanese models. In the experiments, we evaluated the performance of language models. Finally, we discuss the challenges of building Japanese EL models and future works. Therefore, the main contributions of this study are summarized as follows.

- In evaluating the translated WebQSP, the Japanese EL model using the Pointer Network outperforms

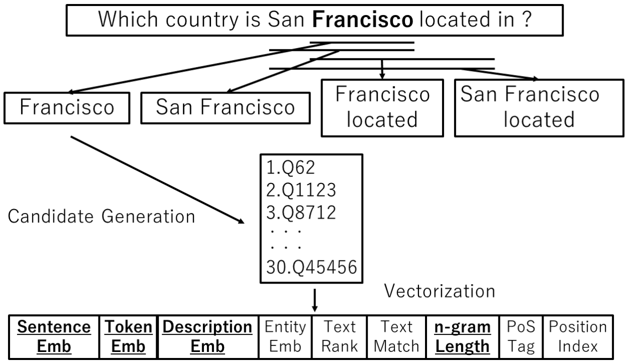

Figure 1: The overview of PNEL

the multilingual EL model by 0.220 points in F1 value.

- An ablation study of the embeddings reveals that it is possible to construct EL models by introducing language-specific embeddings.
- We implement new Japanese EL datasets for Wikidata.

2 RELATED WORK

Studies on EL can be classified as multilingual or a specific language.

Regarding multilingual EL, Botha et al. [6] enabled the model to link to more than 100 languages by adding multilingual settings to the dual encoders. Nicola De Cao et al.[7] predicts the name of the target entity left-to-right, token-by-token in an autoregressive fashion. This model attempted EL for 125 languages, including Japanese, by applying transfer learning and finetuning the model pre-trained on Wikipedia articles in 125 languages. We could not find any other Japanese EL models that could reproduce the experiment, so we used mGENRE in a comparative evaluation.

Regarding EL for a specific language, PNEL [2] for English focused on the size of the input vectors to Pointer Network [17] is not limited. they combined the nine types of embeddings as an input vector to the Pointer Network, and ran EL for Wikidata. Our study built the Japanese EL model based on PNEL. Details of the PNEL are given in Chapter 3. Most EL studies focus on English, and few EL studies focus on languages other than English. Rosales et al. [13] provided results of a study comparing selected entity linking APIs for equivalent documents and sentences in English and Spanish.

3 PNEL

Figure 1 shows an overview of PNEL. PNEL is an EL model based on Pointer Network and has the feature that the size of the input vectors is not restricted. We focus on this flexibility and use PNEL.

3.1 Candidate Generation

The candidate generation process is described in detail in Figure 1. The sentence "Which country is San Francisco located in" is divided into seven tokens: "which", "country", "is", "San", "Francisco", "located" and "in". Then, candidate generation is performed for each token in the following way.

(1) We show the candidate generation process for the term "Francisco" as an illustrative case. Initially, we generate four candidates; "Francisco", "San Francisco", "Francisco located " and "San Francisco located". These candidates represent token sequences containing three or fewer tokens, achieved by concatenating tokens preceding and following the mention of "Francisco".
(2) Performing a label search on the "Francisco" token enables us to retrieve the top 30 candidate entities.
(3) We perform the identical procedure as (2) to obtain the top 30 candidate entities for the tokens "San Francisco", "Francisco located" and "San Francisco located". Consequently, each mention yields a total of 120 candidates.

3.2 Vectorization

After performing the Candidate Generation step, a vector of 1142 dimensions is constructed for each of the 120 candidate entities. These vectors are created by combining nine distinct types of embeddings and inputs to the Pointer Network.

Sentence Embedding Average of word embeddings by fastText [3] of all tokens in the input sentence. (length 300)

Token Embedding Word embeddings in input sentences with fastText. (length 300)

Description Embedding Average of word embeddings by fastText of all tokens in Wikidata descriptions for candidate entities. (length 300)

Entity Embedding Pretrained TransE [5] embeddings [1] built using Wikidata. (length 200)

POS Tags A one-hot vector indicating the token's tag. We use [14]. (length 36)

Text Match Metric Simple ratio, partial ratio, and token sort ratio. These three ratios in numerical expressions range from 0 to 100. In the case of a simple ratio, the following pair of text corresponds to a perfect match: "Elon Musk" and "Elon Musk". In partial ratio, the following pair of text corresponds to a perfect match: "Elon Musk" and "Musk". In the case of the token sort ratio, the following pair of text corresponds to a perfect match: "Elon Musk" and "Musk Elon". (length 3)

Text Rank Search ranking from 0 to 100 of candidate entities. (length 1)

n-gram Length Length of the n-gram from 1 to 3. (length 1)

Position Index Position of the mention in the input sentence. (length 1)

4 PNEL FOR JAPANESE

To construct the Japanese PNEL, we changed four of the nine embeddings in English PNEL. The relevant embeddings are shown as underlined in Figure 1. The four embeddings that have been changed are shown below.

Sentence Embedding Change to the Pre-trained fastText model in Japanese[2].

Token Embedding Change to the Pre-trained fastText model in Japanese, the same one used for Sentence Embedding.

[1] https://torchbiggraph.readthedocs.io/en/latest/pretrained_embeddings.html#wikidata

[2] https://fasttext.cc/docs/en/crawl-vectors

Description Embedding Change to the Pre-trained fastText model in Japanese, the same one used for Sentence Embedding and Token Embedding.

POS Tags Change the tokenization method from the TextBlob[3] library to the Ginza[4] morphological analyser.

Pre-trained fastText is a model trained on Common Crawl and Wikipedia and provides pre-computed vectors for a word. The length of the POS Tags was changed from 36 to 18. The other five embeddings, namely Entity Embedding, Text Match Metric, Text Rank, n-gram Length, and Position Index remained the same.

5 EXPERIMENTS

This experiment starts with a dataset analysis in section 5.1. In section 5.2, an evaluation is conducted comparing the Japanese PNEL with the multilingual EL model [7]. Section 5.3 describes an ablation study aimed at discerning similarities and distinctions between Japanese EL and English EL. For Japanese EL, the selection of candidate entities is constrained to those featuring Japanese labels or descriptions, while for English EL, candidate entities are restricted to English. Notably, EL encompasses two primary aspects: phrase spotting and entity disambiguation tasks. We execute an End-To-End EL evaluation in this experiment while ignoring the two aspects. Finally, in section 5.4, We assess the performance of the language models utilizing four different language models. The material for reproducing the experiments and the appendix of this work are available on github[5].

5.1 Japanese language support for datasets

Following the same approach as Banerjee et al. [2], we use three EL datasets for Wikidata. The datasets are shown below.

LC-QuAD2.0
LC-QuAD2.0 [8] is the most recent dataset of the three. It was built on data from human responses using Amazon Mechanical Turk. It comprises a mixture of complex questions involving more than two entities with the correct answer within a single sentence and simple questions with a single-sentence correct answer.

SimpleQuestions
SimpleQuestions [4] contains only simple questions. It was initially created for Freebase and adapted for Wikidata.

WebQSP
WebQSP[15] has complex questions and simple questions. This dataset contains questions collected from web search logs.

We translate the above datasets into Japanese using the Google Translate API to evaluate the Japanese model.

Table 1 presents the dataset statistics. We selected WebQSP as the dataset for conducting the ablation study on the embedding method based on three considerations: the dataset's size, its minimal susceptibility to accuracy loss when eliminating invalid Japanese data, and the balanced representation of complex and simple questions. We emphasized this balance to ensure the model's applicability in real-world scenarios. In practical applications, cases involving a single correct entity within a sentence and those with multiple valid entities are encountered. Therefore, we used this criterion to evaluate the model without favoring one case over the other, thus avoiding bias.

5.2 Evaluation with the multilingual EL model

In this experiment, we tested the performance of the Japanese PNEL and the multilingual EL model(mGENRE) model [7] on the translated version of WebQSP into Japanese. The multilingual model was tested with and without thresholds, and we used three different threshold patterns. When no threshold was set, the model provided the top five candidates with the highest score. The results of the experiment are shown in Table 2. The F1 value for the Japanese PNEL was 0.673, while for the English PNEL, it was 0.288. The mGENRE model, with a threshold of 0.50, had the highest F1 value. The Japanese PNEL model outperforms the mGENRE model, set with a threshold of 0.50 by 0.220 points in F1 values.

5.3 Ablation study of the embeddings

To provide the Pointer Network with a comprehensive input vector, PNEL combines nine different types of embeddings. To better understand the impact of each embedding on the overall accuracy, we conducted an ablation study on the WebQSP dataset. Like Nicola et al. [7], we used eight of the nine embeddings except for the Position Index. The results are shown in Table 3. The ablation study in English is cited from [7], so only F1 values are displayed. The F1 values for the Japanese PNEL were the lowest when Entity Embedding was omitted, and this result is also seen for the English PNEL. The impact of Text Rank and n-gram Length on the Japanese PNEL was smaller than on the English PNEL.

5.4 Effects of the language models

We utilize fastText to generate three types of embeddings: Sentence Embedding, Description Embedding, and Token Embedding. To study the impact of using different language models, we substitute fastText with other language models such as Wikipedia2Vec [18], chive [9] and WikiEntVec [16] at the vector length of 300. The results are presented in Table 4. We observe that fastText yields the highest F1 score among all the language models used in the experiment.

We experimented with changing the vector length of pre-trained models for Wikipedia2Vec and WikiEntVec. Table 5 displays the results, indicating that precision and recall values at the vector length of 300 are less than 100, and shorter vectors produce higher F1 values.

6 DISCUSSION

Table 3 shows that knowledge graph embedding is the most effective for both the English PNEL and the Japanese PNEL. This is because semantic relations such as entity type and category information are effective in EL, and knowledge graph embedding includes such information. However, we would like to explore the impact of knowledge graph embedding on EL accuracy in detail by using knowledge graph embedding other than TransE. Moreover, Table 3 indicates that Text Rank and n-gram length are ineffective

[3] https://textblob.readthedocs.io/en/dev

[4] https://megagonlabs.github.io/ginza/

[5] https://github.com/ke-lab-it-agu/PNEL-Japanese

Table 1: Number of questions and entities for each dataset

		LC-QuAD2.0		SimpleQuestions		WebQSP	
		train	test	train	test	train	test
original	questions	44,451	5,995	34,374	9,961	3,098	1,638
	questions	63,336	8,327	34,374	9,961	3,378	1,886
except invalid data in English	questions	43,423	5,882	34,212	9,914	3,098	1,638
	entities	61,493	8,131	34,212	9,914	3,378	1,886
except invalid data in Japanese	questions	37,534	5,124	18,741	5,404	3,025	1,597
	entities	52,654	7,022	18,741	5,404	3,283	1,825

Table 2: Evaluation on WebQSP with the multilingual EL model

model	precision	recall	F1
Japanese PNEL	0.830	0.565	**0.673**
English PNEL	0.655	0.185	0.288
mGENRE	0.198	0.664	0.305
mGENRE(0.50)	0.514	0.404	0.453
mGENRE(0.75)	0.351	0.547	0.428
mGENRE(1.00)	0.273	0.617	0.378

Table 3: Ablation study of embeddings on WebQSP

embedding	Japanese			English
	precision	recall	F1	F1
-	0.830	0.565	0.673	0.712
Sentence Embed.	0.753	0.555	0.640	0.554
Token Embed.	0.841	0.520	0.642	0.666
Description Embed.	0.827	0.546	0.658	0.700
Entity Embed.	0.559	0.385	0.456	**0.221**
POS Tags	0.813	0.519	0.633	0.685
Text Rank	0.763	0.532	0.627	0.399
n-gram Length	0.836	0.556	0.668	0.554
Text Match Metric	0.727	0.552	0.628	0.689

Table 4: Evaluation with the pre-trained language models

model	algorithm	precision	recall	F1
fastText	CBOW	0.830	0.565	**0.673**
Wikipedia2Vec	Skip-gram	0.790	0.544	0.645
chive	Skip-gram	0.865	0.518	0.648
WikiEntVec	Skip-gram	0.819	0.531	0.637

Table 5: Evaluation with the pre-trained language models with different length of vectors

model	model	precision	recall	F1
Wikipedia2Vec	300	0.790	0.544	0.645
Wikipedia2Vec	100	0.889	0.563	**0.667**
WikiEntVec	300	0.819	0.521	0.637
WikiEntVec	200	0.775	0.559	0.649
WikiEntVec	100	0.838	0.539	**0.656**
-	-	0.800	0.539	0.644

for the Japanese PNEL. Therefore, we examined individual cases to understand why these embeddings were ineffective.

The highest F1 value in Table 4 was obtained using fastText, possibly because of the differences in the algorithms. We assumed that CBOW-based language models may be more suitable than skip-gram models for EL.

In Table 5, we observe that a shorter vector length of the language model results in a higher F1 value. This is because word embeddings complement knowledge graph embedding, and if the vector size exceeds a specific size, noise may be added to the knowledge graph embedding. The evidence is that knowledge graph embedding is the most effective in Table 3. Therefore, by shortening the length of the word embeddings, we increased the proportion of knowledge graph embeddings, resulting in a higher F1 value. Conversely, the use of word embedding resulted in lower F1 values.

7 CONCLUSION

This study proposes Japanese PNEL as the first Japanese EL model using Pointer Network. The evaluation experiments conducted on the translated WebQSP show that our method outperforms the multilingual model in terms of F1 values. The results of the ablation study indicated that knowledge graph embedding is as effective in Japanese as in English. However, we observed that some embeddings had a relatively minor effect on Japanese PNEL compared to English PNEL. In addition, we found that the model with a shorter word vector length produces a higher F1 value for Japanese PNEL in our language model experiments.

In the future, we plan to investigate the impact of using knowledge graph embeddings other than TransE for Japanese EL and to build an architecture that considers grammatical features in Japanese. We also plan to investigate the effectiveness of this method for languages other than Japanese, such as Chinese and Spanish. Furthermore, we plan to evaluate the performance of the proposed method on other datasets such as LC-QuAD2.0 [8] and Simple Questions [4].

8 ACKNOWLEDGMENTS

This paper is based on results obtained from projects, JPNP20006 and JPNP180013, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This work was partially supported by JSPS KAKENHI Grant Number 23K11221.

REFERENCES

[1] Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. Refined: an efficient zero-shot-capable approach to end-to-end entity linking. *arXiv preprint arXiv:2207.04108*.

[2] Debayan Banerjee, Debanjan Chaudhuri, Mohnish Dubey, and Jens Lehmann. 2020. Pnel: pointer network based end-to-end entity linking over knowledge graphs. In *The Semantic Web – ISWC 2020*, 21–38.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5, 135–146.

[4] Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.

[5] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

[6] Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, (Nov. 2020), 7833–7845.

[7] Nicola De Cao et al. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10, 274–290.

[8] Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019. Lc-quad 2.0: a large dataset for complex question answering over wikidata and dbpedia. In *The Semantic Web – ISWC 2019*, 69–78.

[9] Sorami Hisamoto, Takashi Yamamura, Akihiro Katsuta, Yuto Takebayashi, Kazuma Takaoka, Yoshitaka Uchida, Teruaki Oka, and Masayuki Asahara. 2020. Chive: towards industrial-strength japanese word vector resources–constructing and improving embedding with tokenizer. *IEICE Technical Report; IEICE Tech. Rep.*, 120, 166, 40–45.

[10] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, 782–792.

[11] Yugo Murawaki and Shinsuke mori. 2016. Wicification for scriptio continua. In *Proceedings of 2016 Language Resources and Evaluation Conference*. Association for Computational Linguistics, 1346–1351.

[12] Mikhail Plekhanov, Nora Kassner, Kashyap Popat, Louis Martin, Simone Merello, Borislav Kozlovskii, Frédéric A Dreyer, and Nicola Cancedda. 2023. Multilingual end to end entity linking. *arXiv preprint arXiv:2306.08896*.

[13] Henry Rosales-Méndez, Bárbara Poblete Labra, and Aidan Hogan. 2017. Multilingual entity linking: comparing english and spanish.

[14] Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the penn treebank project.

[15] Daniil Sorokin and Iryna Gurevych. 2018. Mixing context granularities for improved entity linking on question answering data across entity categories. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. (June 2018), 65–75.

[16] Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki, and Kentaro Inui. 2018. A joint neural model for fine-grained named entity classification of wikipedia articles. *IEICE Transactions on Information and Systems*, 101, 1, 73–81.

[17] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in neural information processing systems*, 28.

[18] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: an efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 23–30.

[19] Ikuya Yamada, Koki Washio, Hiroyuki Shindo, and Yuji Matsumoto. 2022. Global entity disambiguation with bert. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3264–3271.