A Pipeline Japanese Entity Linking System with Embedding Features

Shuangshuang Zhou **Koji Matsuda** **Ran Tian** **Naoaki Okazaki** **Kentaro Inui**
Graduate School of Information Sciences, Tohoku University
6-6 Aramaki Aza Aoba, Aobaku, Sendai, Miyagi 980-8579, Japan
{shuang,matsuda,tianran,okazaki,inui}@ecei.tohoku.ac.jp

Abstract

Entity linking (EL) is the task of connecting mentions in texts to entities in a large-scale knowledge base such as Wikipedia. In this paper, we present a pipeline system for Japanese EL which consists of two standard components, namely candidate generation and candidate ranking. We investigate several techniques for each component, using a recently developed Japanese EL corpus. For candidate generation, we find that a concept dictionary using anchor texts of Wikipedia is more effective than methods based on surface similarity. For candidate ranking, we verify that a set of features used in English EL is effective in Japanese EL as well. In addition, by using a corpus that links Japanese mentions to *Japanese* Wikipedia entries, we are able to get rich context information from Japanese Wikipedia articles and benefit mention disambiguation. It was not directly possible with previous EL corpora, which associate mentions to *English* Wikipedia entities. We take this advantage by exploring several embedding models that encode context information of Wikipedia entities, and show that they improve candidate ranking. As a whole, our system achieves 82.27% accuracy, significantly outperforming previous work.

1 Introduction

Entity Linking (EL), also known as wikification or named entity disambiguation, is the task of linking mentions in texts to entities in a large-scale knowledge base such as Wikipedia[1]. EL is useful in many NLP tasks such as information retrieval (Blanco et al., 2015), question answering (Khalid et al., 2008), searching digital libraries (Han et al., 2005), semantic search,[2] coreference resolution (Durrett and Klein, 2014; Hajishirzi et al., 2013), named entity recognition (Durrett and Klein, 2014) and knowledge base population (Suchanek and Weikum, 2013; Dredze et al., 2010).

However, development of Japanese EL has been slow, partly due to the lack of a publicly available Japanese EL corpus. Most previous Japanese EL systems link mentions to English Wikipedia (Furakawa et al., 2014; Nakamura et al., 2015; Hayashi et al., 2014), which might be less informative because there are about 0.44 million articles in Japanese Wikipedia that do not have correspondence in English. Recently, Jargalsaikhan et al. (2016) released a Japanese EL corpus in which mentions are linked to Japanese Wikipedia entries. In this paper, we investigate several techniques for developing a Japanese EL system, and evaluate on this newly released corpus.

An EL system first performs Named Entity Recognition to detect and classify spans of texts which are mentions to certain types of entities. Then, the system links the mentions to entries in Wikipedia. A major challenge here is the mention ambiguity; for example, given the sentence "*The I.B.M. is the world's largest organization dedicated to the art of magic.*", an EL system should associate "*I.B.M*" with the organization "*International Brotherhood of Magicians*", rather than the American technology and consulting company. An or-

[1] https://en.wikipedia.org

[2] https://stics.mpi-inf.mpg.de/

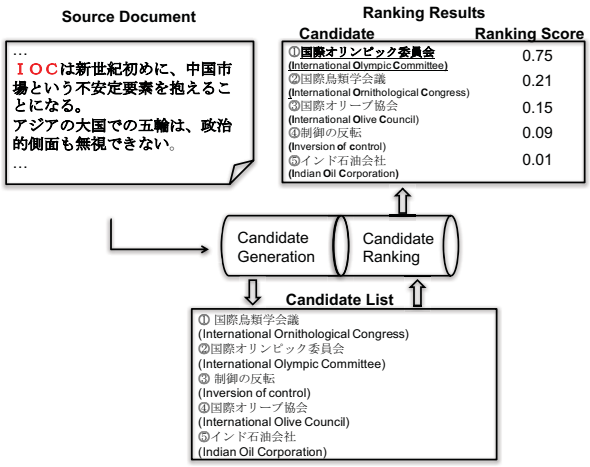

Figure 1: An Entity Linking system generates and ranks a list of candidate entities for the mention "*IOC*".

thodox approach to address this issue is a pipeline of two components, the **candidate generation** component which generates a candidate list of possible entities for each mention, and the **candidate ranking** component which ranks candidates according to multiple features (Figure 1). For candidate generation, another challenge is the variety of mentions. For example, both "*Big Blue*" and "*I.B.M.*" could refer to "*International Business Machines Corporation*".

We investigate several techniques from each component. For candidate generation, string matching between mentions and entity titles has been the main approach, but we find the recall of string matching not satisfactory; instead, a cross-lingual dictionary turns out to be effective in finding correct candidates (Section 3.1, Section 5.3). For candidate ranking, we explore a set of features used in English EL, and find it effective in Japanese EL as well (Section 3.2, Section 5.4). In addition, we apply several embedding models to encode context information of entities in Wikipedia articles, and show that the embeddings are useful features for disambiguating mentions in texts (Section 4, Section 5.4). This technique would not be directly possible in previous Japanese EL systems which link mentions in Japanese texts to English Wikipedia entries, because the embedding models should be trained on articles written in the same language as texts. As a whole, our system achieves 82.27% accuracy and significantly outperforms previous work (Section 5.5).

2 Related Work

English EL is a widely studied topic. There are several public corpora for English EL (Cucerzan, 2007; Yosef et al., 2011), and the TAC-KBP workshop has provided systematical evaluation on EL task in recent years (Ji et al., 2014).

To address mention ambiguity, previous works have explored advanced linguistic features (Bunescu and Pasca, 2006; Dredze et al., 2010; Zhang et al., 2011; Graus et al., 2012; Zhou et al., 2014) and link-based features (Milne and Witten, 2008; Han and Zhao, 2009; Kulkarni et al., 2009; Guo et al., 2011; Ratinov et al., 2011; Hoffart et al., 2011).

Embedding features have been actively used as well. For example, He et al. (2013) use neural networks to compute representations for entities and mentions directly from knowledge base; similarly, Sun et al. (2015) propose to model an entity by combining the sum of surface word vectors and the sum of category word vectors; Blanco et al. (2015) propose mapping entities into word embeddings by using entity descriptions; Hu et al. (2015) build entity hierarchy embedding by learning distance metric of category nodes in Wikipedia; Yang et al. (2014) and Lin et al. (2015) encode relational information by low-dimensional representations.

To counter the variety of mentions, previous English EL systems generate entity candidates by search engine (Dredze et al., 2010; Zhou et al., 2014; Graus et al., 2012), and/or utilize various resources such as Wikipedia disambiguation, Wikipedia redirection, Geonames, *etc.* (Dredze et al., 2010; Zhou et al., 2014).

On the other hand, research on Japanese EL has received less attention. Furakawa et al. (2014) focus on entity linking in academic fields, and link technical terms to English Wikipedia. Nakamura et al. (2015) link keywords in twitter texts to English Wikipedia, aiming at constructing a cross-language topic recognition system. Hayashi et al. (2014) study EL on both English and Japanese texts. In addition, there are several works on linking geopolitical entities in local news articles (Osada et al., 2015; Inoue et al., 2016; Seiya et al., 2015). For candidate generation, most previous Japanese EL systems sim-

ply use surface string matching (Osada et al., 2015; Inoue et al., 2016; Seiya et al., 2015).

3 System Architecture

In this section, we present our pipeline system for Japanese EL. The system takes Named Entity Recognition (NER) as input, and links the named entity mentions to Wikipedia articles as output. For NER, we simply use golden annotations in corpus.

Our system consists of two standard components: candidate generation and candidate ranking (Figure 1). In the candidate generation phase, our system generates a list of Wikipedia articles for each mention in text. For example, given a mention "*IOC*", the candidates which the mention can be linked to include Wikipedia articles titled "国際鳥類学会議 (*International Ornithological Congress*)", "国際オリンピック委員会 (*International Olympic Committee*)", *etc*. Then, in the candidate ranking phase, each Wikipedia article in the candidate list obtains a ranking score from a scoring function, which is constructed via supervised learning on a set of features. We pick the top-1 candidate from the ranking result as system output. For example, in Figure 1, "国際オリンピック委員会 (*International Olympic Committee*)" is output as the referent of "*IOC*". Details of the two components are described below.

3.1 Candidate Generation

If an EL system cannot include correct Wikipedia articles on lists in candidate generation, the next candidate ranking process will be in vain. Previous English EL systems usually generate a candidate list as long as possible. String matching between metion and article titles is a common method for candidate generation.

In this work, we use the simple and efficient `Simstring`[3] tool for calculating similarity and searching similar strings. The tool implements two similarity measures, the cosine similarity and overlap coefficient. We extract all Japanese Wikipedia titles into a database, and use `Simstring` to find all titles with similarity scores larger than a threshold for each mention.

Another approach to candidate generation is the **concept dictionary** (Spitkovsky and Chang, 2012). This approach gathers hyper-links that jump to each Wikipedia article, and regard the surface texts of hyper-links as possible mentions to the article. We call the surface texts of hyper-links **anchor texts**. For example, there are hyper-links in Wikipedia with surface texts "*IOC*", "*I.O.C*" and "*the Olympic Committee*", all jump to the article "国際オリンピック委員会 (*International Olympic Committee*)". Thus, "*I.O.C*" is an anchor text of the article. A concept dictionary is a collection of anchor texts.

[3] http://www.chokkan.org/software/simstring/

3.2 Candidate Ranking

We formulate the candidate ranking problem similar to Bunescu and Pasca (2006) and McNamee et al. (2009). Namely, we construct a scoring function $f(m, e)$ based on features extracted from mention m and candidate Wikipedia article e. We select candidate from a candidate list E, according to the ranking score:

$$\widehat{e} = \arg\max_{e \in E} f(m, e).$$

Therefore, the scoring function $f(m, e)$ should be trained such that the correct Wikipedia article $\widehat{e}$ is linked to the mention m. We use SVM$^{\text{rank}}$ (Joachims, 2006) with linear kernel for training.

3.2.1 Feature Sets

In this section, we describe the features we use to construct the scoring function. These are powerful features used by state-of-the-art English EL systems, combined with several new embedding features. Table 1 shows a complete list. As a running example, we consider the following text snippet (translated from Japanese) surrounding a mention "*IOC*":

> ＩＯＣは新世紀初めに、中国市場という不安定要素を抱えることになる。アジアの大国での五輪は、政治的側面も無視できない。
>
> The **IOC** is facing the elements of instability from the market of China from the beginning of this new century. The Olympics at major Asian nations can never ignore this kind of political aspects.

In which, underlined words are annotated named entities.

Feature Type	Description	Example
String Similarity (S)	string similarity between mention and entity title	the Levenshtein edit-distance between "*IOC*" and "*International Olympic Committee*" is 11
Entity Popularity (P)	distribution of anchor texts in Wikipedia	68% of mention "*IOC*" in Japanese Wikipedia is linked to article "*International Olympic Committee*"
Bag-of-Word (Bw)	BoW similarity between text and Wikipedia article	words {"*face*", "*market*", ...} from text and {"*modern*", "*Olympic*", ...} from Wikipedia article
Bag-of-Entity (Be)	BoE similarity between text and Wikipedia article	entities {"*China*", "*Olympic*", ...} in text and {"*Olympic Games*", ...} in Wikipedia article
Word Vector (WV)	cosine similarity between sums of word vectors	cosine similarity between vector $\mathbf{w}_{face} + \mathbf{w}_{market} + \ldots$ for text and vector $\mathbf{w}_{modern} + \mathbf{w}_{Olympic} + \ldots$ for Wikipedia article
Entity Vector (EV)	cosine similarity between sums of entity vectors	cosine similarity between $\mathbf{e}_{China} + \mathbf{e}_{Olympic} + \ldots$ and $\mathbf{e}_{Olympic\_Games} + \ldots$
Paragraph Vector (PV)	cosine similarity between paragraph vectors	cosine similarity between paragraph vector for text and paragraph vector for Wikipedia article
Entity Category (Cate)	word in text is category of Wikipedia article	Wikipedia article "*International Olympic Committee*" belongs to categories "*Olympic movement*", "*Committees*"
Entity Class (Class)	overlap of Sekine's entity class	mention "*IOC*" in text and Wikipedia entry "*International Olympic Committee*" both labeled *Organization*

Table 1: Features for candidate ranking.

Correspondingly, we show a snippet of the Wikipedia article "国際オリンピック委員会 (*International Olympic Committee*)":

> 国際オリンピック委員会は、近代オリンピックを主催する団体であり、またオリンピックに参加する各種国際スポーツ統括団体を統括する組織である。2009年に国際連合総会オブザーバー資格を得たため国際機関の一つと思われている。

> International Olympic Committee is an organization sponsored by the modern Olympics, also is an organization that oversees the various international sports governing body to participate in the Olympic Games. It is believed to be one of the order to give the General Assembly of the United Nations observer status international organizations in 2009.

In which, underlined words are anchor texts (i.e. hyper-links).

We consider the following features.

String Similarity This type of features measures the string similarity between mentions and the titles of Wikipedia articles. We use several similarity measures explored in previous work (Graus et al., 2012; Dietz and Dalton, 2012), such as the Levenshtein edit distance and Jaccard coefficient score.

Entity Popularity This is the probability $p(e|m)$ of an anchor text m linking to a Wikipedia article e. The probability is estimated as:

$$p(e|m) = \frac{\text{\# times of } m \text{ jumping to } e}{\text{\# occurrence of anchor text } m}.$$

As discussed in Milne and Witten (2008), this probability reflects the "commonness" or "popularity" of a Wikipedia article.

Bag-of-Word Similarity This feature measures the similarity between texts surrounding the mention and the contents of the Wikipedia article. For example, we assess the similarity between the set of words {"*face*", "*market*", ...} taken from text, and the set of words {"*modern*", "*Olympic*", ...} taken from Wikipedia article. We consider several similarity measures such as cosine similarity of TF-IDF weights (Zheng et al., 2010) and Jaccard coefficient (Dietz and Dalton, 2012).

Bag-of-Entity Similarity This is similar to Bag-of-Word Similarity, except that we only take named entities from text and anchor texts from Wikipedia articles. For example, we assess the similarity between the set of entities {"*China*", "*Olympic*", ...} taken from text, and the set of anchor texts {"*Olympic Games*", ...} taken from Wikipedia article.

Embedding Similarity We construct vectors for texts and Wikipedia articles, and assess cosine similarity between the vectors. This feature also measures the similarity between texts and Wikipedia contents. We consider three types of vectors, namely the word vector (WV), entity vector (EV), and paragraph vector (PV). Details of the embedding models are described in Section 4.

Entity Category This feature counts how many words in category names of a Wikipedia article also appear in text. For example, the Wikipedia article "*International Olympic Committee*" belongs to categories "*Olympic movement*", "*Committees*", *etc.*, and some words in the category names, such as "*Olympic*", also appear in text. This feature reflects such overlaps.

Entity Class The corpus (Jargalsaikhan et al., 2016) we use in this work has annotated each named entity with a fine-grained entity class label, called Sekine's entity class (Sekine et al., 2002). On the other hand, Suzuki et al. (2016) released a system which automatically label Wikipedia articles with Sekine's entity classes. We use this system and assess overlap between the two entity class labels. For example, the Wikipedia article "*International Olympic Committee*" is assigned the class label "*Sports Organization Other*", whereas the mention"*IOC*" is annotated as "*International Organization*"; both of them are organizations. It has been shown that finer-grained entity class is useful for English EL (Ling and Weld, 2012; Ling et al., 2015).

4 Embedding Models

In this section, we describe the embedding models we use to construct vectors for texts and Wikipedia articles.

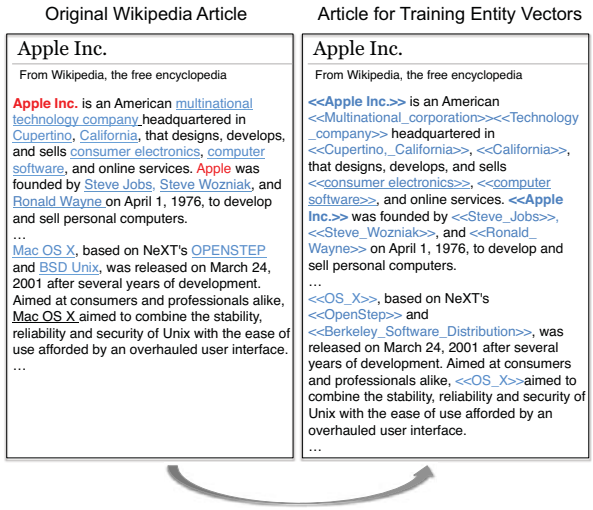

Figure 2: Training entity vectors from anchor texts.

4.1 Word Vector

We apply the `word2vec`[4] tool to Japanese Wikipedia for training word vectors. Then, we take sums of the word vectors to obtain document vectors.

4.2 Entity Vector

The Skip-gram model (Mikolov et al., 2013b) implemented by `word2vec` learns vectors by predicting context words from targets. We use this model to train vectors for Wikipedia articles, by regarding each anchor text as a target of the referent article, and words surrounding the anchor text as context. For example, in Figure 2 we replace all anchor texts with their referent articles (e.g. converting the hyper-link "*Mac OS X*" to $\langle\langle OS\_X \rangle\rangle$, representing the Wikipedia article "*OS X*"), and train vectors for the referent articles according to the converted document.

4.3 Paragraph Vector

The paragraph vector (Le and Mikolov, 2014) is a powerful unsupervised method of learning representations of arbitrary lengths of texts and has the advantages of simplicity and versatility. We use the Distributed Memory Model of Paragraph Vectors model to train paragraph vectors of texts and Wikipedia articles. The model is an extension of the

[4]https://code.google.com/p/word2vec/

CBoW model (Mikolov et al., 2013a) implemented in `word2vec`.

5 Experiments

In this section, we first introduce the evaluation data set. Then we evaluate the performance of candidate generation and the performance of each feature set on candidate ranking. Finally, we compare our system with the previous work (Jargalsaikhan et al., 2016).

5.1 Data set

We use a new released Japanese Wikification corpus (Jargalsaikhan et al., 2016), which consists of 340 newspaper articles from Balanced Corpus of Contemporary Written Japanese (BCCWJ).[5] Mentions in each document are annotated with fine-grained named entity class labels that are defined by Sekine Extended Named Entity Hierarchy (Sekine et al., 2002).[6] In this corpus, 19,121 mentions are linked to Wikipedia while 6,554 mentions do not reference Wikipedia articles. 7,118 distinct mentions were linked to 6,008 distinct entities totally. Because the corpus was built with recognized named entities, we omit the step of mention detection.

Since mentions are scattered in texts of the original corpus, in order to facilitate the system processing, we generate a single document that contains the composite of all mentions. Our new data set contains all information of mentions of which the format refers to the TAC-KBP data set. An example is shown in Figure 3. We obtain the information of a mention including mention ID, document ID, mention name, begin offset, end offset, entity class, entity linking mark, unique Wikipedia ID and unique Wikipedia title.

5.2 Experimental Setup

We utilize 2016.3.5 Japanese Wikipedia dump as the referent Knowledge base. We tokenize and remove punctuations in documents by using a Japapense part-of-speech and morphological analyzer, Mecab.[7] We learn word embeddings, entity embeddings and paragraph vectors on this processed corpus. The

```
<mention id="PN1a_00002_T38">
<name>佐藤秀夫(Sato Hiteo)</name>
<docid>PN1a_00002</docid>
<beg>2687</beg>
<end>2691</end>
<entity class>Person</entity class>
<entity linking mark>A</entity linking mark>
<wikipedia id>2617934</wikipedia id>
<wikipedai title>佐藤秀夫(Sato
Hiteo)</wikipedia title>
</mention>
```

Figure 3: A mention snippet in data set.

word and entity vectors are learned by setting the dimensions d to 200, the size of context window c to 10 and the negative samples to 5. Meanwhile, the paragraph vectors are learned by setting the dimensions d to 400, the size of context window c to 5 and the negative samples to 5.

5.3 Evaluation of Candidate Generation

We evaluate our candidate generation methods on all mentions in the corpus (Jargalsaikhan et al., 2016). We normalize mention surfaces to eliminate the effect of half-width characters or full-width characters in the preprocessing step.

We compare cosine similarity and overlap coefficient with threshold of 0.7 and 0.9 respectively. We look up the concept dictionary with the mention and we can obtain Wikipedia articles from the results of entries. Table 2 shows the results of recall and average length of candidate lists. Here, recall means the percentage of mentions that have the gold entity in the candidate list. Moreover, we also compare the candidate list length because the more counts of candidates we have, the more time will be spend on candidate ranking.

According to the results in Table 2, we find that our concept dictionary based on anchor texts is suitable for the need of high-recall (91.98%) and short length (17.58). Moreover, we extend family names or given names of person to full names before searching on the concept dictionary, which will enhance the correct rate. After this extending step, we achieved the recall of 94.14% and the average number of candidates per list is 17.79.

[5] http://pj.ninjal.ac.jp/corpus_center/bccwj

[6] https://sites.google.com/site/extendednamedentityhierarchy/

[7] http://taku910.github.io/mecab/

Methods	Recall	AveLen
cosine(Threshold=0.9)	74.49%	1.58
overlap(Threshold=0.9)	66.68%	736.4
cosine(Threshod=0.7)	87.47%	27.12
overlap(Threshold=0.7)	68.01%	1750
anchor texts	91.98%	17.58
anchor texts (+extended)	94.14%	17.79

Table 2: Performance of candidate generation approaches on NonNIL mentions.

5.4 Feature Study

We conducted the feature study on each feature set by a 5-fold cross validation. We applied experiments on NonNILs, entities that exist in the Wikipedia. We begin with the string similarity feature set, added various features to it incrementally and reported their impact on performance.

From the results of Table 3, we found that our system obtained the performance with approximately 3 percents higher than previous work by only using string similarity features. Adding popularity features slightly further improved the performance.

We observed significant improvement when adding Bag-of-words features. However, only adding Bag-of-entities features led the performance to drop by about 9 percents. Adding both Bag-of-words and Bag-of-entities together, the system performance is improved to 84.88%.

Moreover, adding the features of fine-grained entity class is better than adding the category features. Therefore, we remove the category feature in the remaining experiments.

In addition, our system had slightly improved by adding entity embedding features. Here, features of entity vectors (EV) is more effective than features of word vectors (WV) by the accuracy of 0.64%. We also found that only using features of entity vectors (EV) is better using both word vectors (WV) and entity vectors (EV). The best performance of our system reached to 86.68% after adding features of paragraph vectors (PV).

5.5 System Performance

We made a 5-fold cross validation and calculated the average accuracy of each fold. Although we get the top-1 Wikipedia article from the ranking results, we need to determine that the mention in the text is a NonNIL or a NIL. NonNILs are entities that exist in the KB (Wikipedia) while NILs are entities that do not exist in the KB (Wikipedia).

In NIL labeling, we use two rules to make decisions. First, the mention will be labeled with NIL when there is no Wikipedia article for it. Second, the mention will be labeled with NIL when the ranking score of the top 1 candidate of the mention is below a threshold (heuristically set to 2.9).

Table 4 shows the accuracy of our system as well as a unsupervised method (Jargalsaikhan et al., 2016). Their method relies on the popularity of entities in the anchor texts of the mention, which is the same with our *Entity Popularity* feature. They also estimate probability distributions conditioned on a mention and its fine-grained semantic classes. We compared system performance of NILs and NonNILs while there is no comparison in the previous work (Jargalsaikhan et al., 2016). Our proposed system achieved a 82.27% accuracy across the 5-folds and outperform the previous unsupervised method by significant margins.

5.6 Error Analysis

For our candidate generation method, we found that some failure cases are caused by transliterating katakana from other languages. Since the abbreviation rules of Japanese are different from English, some failure cases are caused by lacking of resources to obtain specific abbreviations of Japanese characters.

Moreover, we found that exactly surface matching and high popularity have strong bias effects on incorrect entities. For example, a mention "*Japan*" may refer to the entity "Japan Television Network Corporation" in the sentence "There is a logo 'Old men can have beautiful life' in Beauty 7 (*Japan* 10:00PM)". However, the incorrect entity "*Japan* (Country)" is linked because of the bias effects. Furthermore, lacking of description words in Wikipedia is also a problem for our context based method.

Finally, we utilized the simple rules for NIL labeling instead of learning the characters of NILs. Talbe 4 shows our system performance on NILs is far from that of NonNILs. The weak NILs performance slightly affected the whole system performance because the counts of NonNILs is three times

Feature sets	Accuracy
Jargalsaikhan et al. (2016) Popularity	53.31%
StringSim (S)	56.13%
S+Popularity (P)	61.87%
S+P+Bag-of-words (Bw)	84.48%
S+P+Bag-of-entities (Be)	75.26%
S+P+Bw+Be	84.88%
S+P+Bw+Be+Entity Category (Cate)	84.77%
S+P+Bw+Be+Entity Class (Class)	85.54%
S+P+Bw+Be+Cate+Class	85.37%
S+P+Bw+Be+Class+Word Vectors (WV)	85.58%
S+P+Bw+Be+Class+Entity Vectors (EV)	86.22%
S+P+Bw+Be+Class+WV+EV	85.79%
S+P+Bw+Be+Class+EV+Paragraph Vectors (PV)	86.68%

Table 3: Performance on NonNILs by incremental feature study.

Methods	Acc(NonNILs)	Acc(NILs)	Acc(All)
Our system	86.95%	68.80%	82.27%
Jargalsaikhan et al. (2016) Popularity	–	–	53.31%
Jargalsaikhan et al. (2016) Popularity + Class	–	–	53.26%

Table 4: Comparing the system performance of the proposed method with an unsupervised method.

of NILs counts.

6 Conclusions and Future Work

In this paper, we constructed a pipeline Japanese EL system that consists of two standard components, candidate generation and candidate ranking. We build a concept dictionary to generate referent Wikipedia articles for Japanese mentions. Comparing with the methods based on surface similarity, the concept dictionary extracted from Wikipedia was verified more effective on generating candidate lists with high-recall and short length.

Moreover, we verified that the effectiveness of several feature sets on Japanese EL that have been used in English EL . We jointly learned a new entity representation model and improved the system performance by adding features based on the learned entity embeddings. We verified that word embeddings and paragraph vectors also effectively improve the system performance. All in all, our system overcome the previous work on the same data set with significant margins.

In future work, we plan to use the cross-lingual information retrieval technology to solve the transliteration problems between Japanese and English. We also consider developing methods to solve the problems of matching abbreviation mentions to Wikipedia articles on Japanese. Moreover, we intend to improve our system by leveraging advanced context embedding methods instead of using the sum of vectors, such as CNN (Convolutional Neural network), LTSM (Long Short Term Memory), etc.

In addition, we will connect mention detection component to our current system and construct an end-to-end Japanese EL system. Finally, we expect the effectiveness of our Japanese EL system on other NLP task, e.g. knowledge base population, question answering, etc.

References

Roi Blanco, Giuseppe Ottaviano, and Edgar Meij. 2015. Fast and space-efficient entity linking for queries. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 179–188. ACM.

Razvan C Bunescu and Marius Pasca. 2006. Using ency-

clopedic knowledge for named entity disambiguation. In *Proceedings of EACL*, volume 6, pages 9–16.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of EMNLP-CoNLL*, volume 7, pages 708–716.

Laura Dietz and Jeffrey Dalton. 2012. A cross document neighborhood expansion: Umass at tac kbp 2012 entity linking. In *Proceedings of Text Analysis Conference (TAC)*.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 277–285.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Proceedings of TACL*, 2:477–490.

Tatsuya Furakawa, Takeshi Sagara, and Akiko Aizawa. 2014. Semantic disambiguation for cross-lingual entity linking (in japanese). *Journal of Japan society of Information and Knowledge*, 24(2):172–177.

David Graus, Tom Kenter, Marc Bron, Edgar Meij, M Rijke, et al. 2012. Context-based entity linking-university of amsterdam at tac 2012.

Yuhang Guo, Guohua Tang, Wanxiang Che, Ting Liu, and Sheng Li. 2011. Hit approaches to entity linking at tac 2011. In *Proceedings of the Fourth Text Analysis Conference (TAC 2011)*. Citeseer.

Hannaneh Hajishirzi, Leila Zilles, Daniel S Weld, and Luke S Zettlemoyer. 2013. Joint coreference resolution and named-entity linking with multi-pass sieves. In *Proceedings of EMNLP*, pages 289–299.

Xianpei Han and Jun Zhao. 2009. Named entity disambiguation by leveraging wikipedia semantic knowledge. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 215–224. ACM.

Hui Han, Hongyuan Zha, and C Lee Giles. 2005. Name disambiguation in author citations using a k-way spectral clustering method. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on*, pages 334–343. IEEE.

Yoshihiko Hayashi, Kenji Yamakuchi, Masaaki Nagata, and Takaaki Tanaka. 2014. Improving wikification of bitexts by completing cross-lingual information (in japanese). In *Proceedings of The 28th Annual Conference of the Japanese Society for Artificial Intelligence*, pages 1A2–2.

Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. 2013. Learning entity representation for entity disambiguation. In *Proceedings of ACL*, pages 30–34.

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 EMNLP*, pages 782–792.

Zhiting Hu, Poyao Huang, Yuntian Deng, Yingkai Gao, and Eric P Xing. 2015. Entity hierarchy embedding. In *Proceedings of ACL-IJCNLP*, volume 1, pages 1292–1300.

Tatsukuni Inoue, Keigo Suenaga, Nagata Seiya, and Kenji Tateishi. 2016. Tagging geopolitical information on news article by using entity linking (in japanese). In *Proceedings of the Twenty-second Annual Meeting of the Association for Natural Language Processing*.

Davaajav Jargalsaikhan, Naoaki Okazaki, Koji Matsuda, and Kentaro Inui. 2016. Building a corpus for japanese wikificaiton with fine-grained entity classes. In *ACL student research workshop. to appear*.

Heng Ji, Joel Nothman, and Ben Hachey. 2014. Overview of tac-kbp2014 entity discovery and linking tasks. In *Proceedings of Text Analysis Conference (TAC2014)*.

Thorsten Joachims. 2006. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226. ACM.

Mahboob Alam Khalid, Valentin Jijkoun, and Maarten De Rijke. 2008. The impact of named entity normalization on information retrieval for question answering. In *Proceedings of Advances in Information Retrieval*, pages 705–710. Springer.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–466. ACM.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of AAAI*, pages 2181–2187.

Xiao Ling and Daniel S Weld. 2012. Fine-grained entity recognition. In *Proceedings of AAAI*.

Xiao Ling, Sameer Singh, and Daniel S Weld. 2015. Design challenges for entity linking. *Proceedings of TACL*, 3:315–328.

Paul McNamee, Mark Dredze, Adam Gerber, Nikesh Garera, Tim Finin, James Mayfield, Christine Piatko, Delip Rao, David Yarowsky, and Markus Dreyer.

2009. Hltcoe approaches to knowledge base population at tac 2009. In *Proceedings of Text Analysis Conference (TAC)*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in neural information processing systems*, pages 3111–3119.

David Milne and Ian H Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM.

Tatsuya Nakamura, Masumi Shirakawa, Takahiro Hara, and Shojiro Nishio. 2015. An entity linking method for closs-lingual topic extraction from social media (in japanese). In *Proceedings of DEIM Forum 2015*, pages A3–1.

Seiya Osada, Keigo Suenaga, Yoshizumi Shogo, Kazumasa Shoji, Tsuneharu Yoshida, and Yasuaki Hashimoto. 2015. Assigning geographical point 559 information for document via entity linking (in japanese). In *Proceedings of the Twenty-first Annual Meeting of the Association for Natural Language Processing*, pages A4–4.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of ACL*, pages 1375–1384. Association for Computational Linguistics.

Nagata Seiya, Keigo Suenaga, Yoshizumi Shogo, Kazumasa Shoji, Yoshida ToruHaru, and Hashimoto KyoAkira. 2015. Application of geopolitical entity linking on documents (in japanese). In *Proceedings of the Twenty-first Annual Meeting of the Association for Natural Language Processing*.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In *Proceedings of LREC*.

Valentin I Spitkovsky and Angel X Chang. 2012. A cross-lingual dictionary for english wikipedia concepts. In *Proceedings of LREC*, pages 3168–3175.

Fabian Suchanek and Gerhard Weikum. 2013. Knowledge harvesting from text and web sources. In *Proceedings of Data Engineering (ICDE)*, pages 1250–1253. IEEE.

Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. 2015. Modeling mention, context and entity with neural networks for entity disambiguation. In *Proceedings of IJCAI*, pages 1333–1339.

Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki, and Kentaro Inui. 2016. Multi-label classification of wikipedia articles into fine-grained named entity types (in japanese). In *Proceedings of the Twenty-second Annual Meeting of the Association for Natural Language Processing*.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Mohamed Amir Yosef, Johannes Hoffart, Ilaria Bordino, Marc Spaniol, and Gerhard Weikum. 2011. Aida: An online tool for accurate disambiguation of named entities in text and tables. *Proceedings of the VLDB Endowment*, 4(12):1450–1453.

Wei Zhang, Yan Chuan Sim, Jian Su, and Chew Lim Tan. 2011. Entity linking with effective acronym expansion, instance selection, and topic modeling. In *Proceedings of IJCAI*, volume 2011, pages 1909–1914.

Zhicheng Zheng, Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010. Learning to link entities with knowledge base. In *Proceedings of NAACL*, pages 483–491. Association for Computational Linguistics.

Shuangshuang Zhou, Canasai Kruengkrai, Naoaki Okazaki, and Kentaro Inui. 2014. Exploring linguistic features for named entity disambiguation. *International Journal of Computational Linguistics and Applications*, 5(2):49.