Cross-language Article Linking using Cross-Encyclopedia Entity Embedding

Chun-Kai Wu
Department of Computer Science
National Tsinghua University
Hsinchu, Taiwan
s106062851@m106.nthu.edu.tw

Richard Tzong-Han Tsai*
Department of CSIE
National Central University
Chungli, Taiwan
thtsai@csie.ncu.edu.tw

Abstract

Cross-language article linking (CLAL) is the task of finding corresponding article pairs of different languages across encyclopedias. This task is a difficult disambiguation problem in which one article must be selected among several candidate articles with similar titles and contents. Existing works focus on engineering text-based or link-based features for this task, which is a time-consuming job, and some of these features are only applicable within the same encyclopedia. In this paper, we address these problems by proposing cross-encyclopedia entity embedding. Unlike other works, our proposed method does not rely on known cross-language pairs. We apply our method to CLAL between English Wikipedia and Chinese Baidu Baike. Our features improve performance relative to the baseline by 29.62%. Tested 30 times, our system achieved an average improvement of 2.76% over the current best system (26.86% over baseline), a statistically significant result.

1 Introduction

Online encyclopedias now make vast amounts of information and human knowledge available to internet users around the world in various languages. However, information resources are not evenly distributed across all languages. To facilitate international knowledge sharing, the task of cross-language article linking (CLAL) aims to create links between encyclopedia articles in different languages that describe the same content. CLAL has been applied in several fields such as named entity translation (Lee and Hwang, 2013), cross-language information retrieval (Nguyen et al., 2009), and multilingual knowledge base creation (Lehmann et al., 2015).

Much CLAL research has been carried out on Wikipedia, one of the largest multilingual online encyclopedias. Articles in Wikipedia are partly structured, usually containing a title, table of contents, main context, related images, media files, categories, and infoboxes (structured metadata tables). Wikipedia articles may also have inter-language links to corresponding articles in other language versions. However, these links are manually created, so some articles lack inter-language links. Several approaches (Sorg and Cimiano, 2008; Oh et al., 2008; Bennacer et al., 2015) have been proposed to automatically generate inter-language links between different language versions of Wikipedia. The main challenge in this task is the distinction of ambiguous candidate articles with similar titles or contents. Most previous work on CLAL has used hand-crafted features for each encyclopedia, which is unscalable.

In this paper, we propose a method to learn cross-encyclopedia entity embedding (CEEE) on English Wikipedia and Chinese Baidu Baike. Every article (entity) in the two encyclopedias is represented by an embedding so that corresponding articles are placed closer together in the vector space. To acquire the training data without human annotation, we also offer a way to discover article correspondence among different encyclopedias. A set of evaluations carried out on the CLAL task between English Wikipedia and Chinese Baidu Baike show that our embedding method improves performance relative to the baseline by 29.62%, which represents a statistically significant improvement over the current best system (26.86% relative improvement). To the best of our knowledge, this is the first work to learn cross-encyclopedia bilingual entity embedding without relying on existing manually labeled cross-language links.

*corresponding author

Proceedings of NAACL-HLT 2018, pages 334–339
New Orleans, Louisiana, June 1 - 6, 2018.

2 Related Work

2.1 Cross-language article linking

Previous Wikipedia-only CLAL studies have relied on existing inter-language links and have focused on using text-based features or link-based features in classifiers. For example, Wang et al.'s (2012) cross-language link similarity work can only be used for linking cross-language Wikipedia articles. Sorg and Cimiano (2008) assumed that given a cross-language pair a and b, the articles linked to a in encyclopedia A are more likely to be linked to the articles linked to b in encyclopedia B. They then designed link-based features on that assumption. Oh et al. (2008) relied on text-based features. They first constructed a translation dictionary from cross-language links. They then translated English terms into Japanese and designed features based on edit distance. Bennacer et al. (2015) used BabelNet (Navigli and Ponzetto, 2012) to select candidate articles with similar semantics in the target language. They then ranked the candidates based on cross-language link similarity. Unlike the above studies, we aim to carry out CLAL between different encyclopedia platforms, English and Baidu Baike; therefore we cannot use existing inter-language links.

Because much of the previous Wikipedia-only CLAL research cannot be directly applied to cross-platform linking, Wang et al. (2014) developed an SVM-based approach with content-similarity-based features to link articles in Wikipedia and Baidu Baike. In this work, they relied on Google Translate to translate Chinese terms into English. They then designed title-based and content-based features based on both the English and translated Chinese articles.

2.2 Entity Embedding

With recent work on word embedding, there has also been more interest in learning entity embedding. Hu et al. (2015) modeled Wikipedia's category structure in their entity embedding. Li et al. (2016) extended Hu et al.'s (2015) work to include category embedding in addition to entity embedding. They further extended the model by integrating category structure to capture meaningful semantic relationships between entities and categories. Yamada et al. (2016) learned joint embedding for words and entities. Tsai and Roth (2016) proposed a way to learn multilingual embedding of words and entities. They first learned monolingual entity and word embedding from the Wikipedia corpus with the skip-gram word embedding model by replacing entity mentions with special symbols. They then used canonical correlation analysis (CCA) to project different embeddings on the same space, learning the CCA model using the existing Wikipedia cross-language links as the training data.

3 Methods

Given an article from a knowledge base (KB), CLAL aims to find the article's corresponding article in another KB of a different language. Corresponding articles are defined as articles describing the same entity in different languages. We base our CLAL system on Wang et al.'s (2014) work because theirs is the only previous CLAL system designed for a cross-encyclopedia setting.

Following Wang et al.'s (2014) example, we also divide CLAL into two stages: candidate selection and candidate ranking. The candidates for each Wikipedia article are selected with the Lucene search engine, and the queries and documents are translated with the Google Translate API. We then train an SVM classifier with the same features described in Wang et al.'s (2014) paper. The given English Wikipedia article and a candidate Baidu article are denoted as w and b. Wang et al.'s (2014) features are as follows:

- BM25: w's title is translated into Chinese and then used as a query to retrieve articles from Baidu Baike with the Lucene search engine. The returned BM25 score corresponding to b is treated as the value of b's BM25 feature.

- Hypernym translation (HT): Supposing the given English title is e and that e's hypernym is h, this feature is defined as the log frequency of h's Chinese translation in the candidate Chinese article.

- English title occurrence (ETO): Whether or not w's title appears in the first sentence of b is regarded as the value of b's ETO feature.

After replicating Wang et al.'s (2014) system, we add our proposed cross-encyclopedia entity embedding (CEEE) feature, the construction of which is detailed in the following sections.

3.1 Cross-Encyclopedia Entity Embedding Model

Similar to (Mikolov et al., 2013), our model learns the entity representation that are useful for predicting the target entity given the context entity. Within an online encyclopedia, each entity is linked with one or more other entities by hyperlinks. For example, the "Food" article in English Wikipedia is linked with the "Plant" article. We treat every article as context entity and the hyper-linked article in a context entity as target entity. Given a set of target-context entity pairs $E_{\mathcal{ST}} = \{(t, c)\}$ where every context entity c comes from the encyclopedia $\mathcal{S}$ and every target entity t comes from the encyclopedia $\mathcal{T}$, we learn the embeddings of entities by maximizing the training objective:

$$\mathcal{L}_{\mathcal{ST}} = \frac{1}{|E_{\mathcal{ST}}|} \Sigma_{(t,c)\in E_{\mathcal{ST}}} log P(t|c). \quad (1)$$

The probability of a target entity given a certain context entity is defined with the softmax function to represent the probability distribution over the entity space ε of the online encyclopedia which the target entity t is residing in:

$$P(t|c) = \frac{exp(v_t \odot v_c)}{\Sigma_{e\in\varepsilon} exp(v_e \odot v_c)}, \quad (2)$$

where $v_t, v_c \subset \mathbb{R}^d$ is the embedding of an entity, d is the size of the embedding and $\odot$ is the dot product operation. Using the link structure of Wikipedia and Baidu Baike, we have compiled two sets of entity pairs, $E_{\mathcal{WW}}$ and $E_{\mathcal{BB}}$, for training Wikipedia and Baidu entity embeddings, respectively.

3.2 Training Data Compilation for Cross-Encyclopedia Entity Embedding

To acquire the training data for cross-encyclopedia bilingual entity embeddings, we first translate all categories of Baidu Baike into English and then collect a set of common category labels between English Wikipedia and Baidu Baike by exact string match. There are 6,297 common categories between Wikipedia and Baidu. If there is no common category then the category is labeled as null. Next, we compile a set of entity pairs from Wikipedia and Baidu articles that share at least 1 common category label. The intuition behind the proposed method is that semantically correlated entities tend to share common category labels, and so do cross-lingual entities.

We then compile another two sets of entity pairs, $E_{\mathcal{WB}}$ and $E_{\mathcal{BW}}$ to learn meaningful cross-encyclopedia entity embeddings. The entity pairs in both sets are identical, except the roles are changed. Specifically, in $E_{\mathcal{BW}}$ the Wikipedia entity is the target while the Baidu entity is the context, and in $E_{\mathcal{WB}}$ the roles are reversed.

3.3 Learning Cross-Encyclopedia Entity Embedding

Since there are millions of entities in both Wikipedia and Baidu, we adopt negative sampling to speed up the training process. We set the negative sample size to 100 during training. We further filter out entities that are only linked to 9 or fewer other entities. Given the two embedding matrices $m^{\mathcal{W}} \subset \mathbb{R}^{|\mathcal{W}|\times d}$ and $m^{\mathcal{B}} \subset \mathbb{R}^{|\mathcal{B}|\times d}$, corresponding to Wikipedia and Baidu, we train the CEEE model with the following 4 tasks: (1) to predict any Baidu article given a Wikipedia article as context by optimizing $\mathcal{L}_{\mathcal{WB}}$, (2) to predict any Wikipedia article given a Baidu article as context by optimizing $\mathcal{L}_{\mathcal{BW}}$, (3) to predict any Wikipedia article given another Wikipedia article as context by optimizing $\mathcal{L}_{\mathcal{WW}}$, and (4) to predict any Baidu article given another Baidu article as context by optimizing $\mathcal{L}_{\mathcal{BB}}$. During task (3), only $m^{\mathcal{W}}$ is updated, and during task (4), only $m^{\mathcal{B}}$ is updated. Every task iterates through its corresponding set of entity pairs. The four tasks repeat 50 times each. The embeddings are updated by stochastic gradient descent with a batch size of 1280 entity pairs. The learning rate is set to 0.1, and entity embeddings are randomly initialized. We also normalize the embeddings to the unit vector every 10 batches during training as Xing et al. (2015) did to improve entity similarity measurement. We set the embedding size d to 100 for the following experiments.

After training, the learned embedding matrices are ready to be used. The similarity score of a Wikipedia entity and a Baidu entity is obtained by calculating the cosine value of their corresponding vectors in the learned embedding matrices. Supposing the embedding vectors corresponding to the English Wikipedia article and the Baidu article are v_w and v_b, the feature value is defined as follows:

$$\begin{cases} \frac{v_w \cdot v_b}{|v_w||v_b|} & \textit{if both } v_w \textit{ and } v_b \textit{ are available} \\ -1 & \textit{otherwise} \end{cases} \quad (3)$$

Table 1: Cross-language article linking performance the first dataset. "MRR (adding CEEE)" column shows the MRR and performance gain. The last column shows t-values. The p-value of every configuration is less than 0.001. Hence, we conclude that CEEE is effective in improving performance.

Configuration	MRR	MRR (adding CEEE)	
BM25	.6146	.7418(**+.1272**)	-29.6
BM25 + HT	.7457	.7944(+.0487)	-16.5
BM25 + ETO	.6469	.7604(+.1135)	-34.1
BM25 + HT + ETO	**.7542**	**.7967**(+.0425)	-14.9

Table 2: Cross-language article linking performance of the second dataset. "MRR (adding CEEE)" column shows the MRR and performance gain. The last column shows t-values. The p-value of every configuration is less than 0.001.

Configuration	MRR	MRR (adding CEEE)	
BM25	.6093	.6762(**+.0669**)	-11.3
BM25 + HT	**.6611**	.6918(+.0307)	-9.1
BM25 + ETO	.6466	.6933(+.0467)	-13.8
BM25 + HT + ETO	.6606	**.6989**(+.0383)	-13.5

Table 3: Performance comparison of CLAL methods. The mean MRR of our system is significantly greater than that of Wang et al.'s (2014), t-value $= -7.6$ and $= -4.5$, respectively. The p-values are less than 0.001 for both datasets.

Dataset	System	MRR	Relative Improvement
1st	BM25 (Baseline)	.6146	
	Our system	**.7967**	**29.62%**
	Wang et al. 2014	.7797	26.86%
2nd	BM25 (Baseline)	.6093	
	Our system	**.6989**	**14.70%**
	Wang et al. 2014	.6874	12.81%

4 Experiments

Following the same procedure used in (Wang et al., 2014), we downloaded the entire English Wikipedia dump and obtained 4M entities and 0.9M categories. We also crawled 6M Baidu Baike articles, which contain 50 thousand distinct category labels. Within Wikipedia and Baidu, there are 68M training pairs for Wikipedia and 20M for Baidu. After matching common categories, we extracted 54M bilingual entity pairs. To generate the gold standard evaluation sets of correct English and Chinese article pairs, we automatically collect English-Chinese inter-language links from Wikipedia. For pairs that have both English and Chinese articles, the Chinese article title is regarded as the translation of the English one. Next, we check if there is a Chinese article in Baidu Baike with exactly the same title as the one in Chinese Wikipedia. If so, the corresponding English Wikipedia article and the Baidu Baike article are paired in the gold standard. We create two different datasets for our experiments. For the first dataset, we select the top 500 English-Chinese article pairs with the highest page view counts in Baidu Baike. This set represents the articles people in China are most interested in. The second dataset is based on random selection. We first randomly select 3500 Wikipedia articles and link them to Baidu Baike articles using English-Chinese Wikipedia inter-links and redirect pages. To eliminate rarely-viewed articles, the 3500 English-Chinese article pairs are sorted in decreasing order based on click count statistics of the article in English Wikipedia in 2012. The top 1000 English-Baidu article pairs are retained as the second dataset. For statistical generality, the data set is randomly split 4:1 (training:test) 30 times. The final evaluation results are calculated as the mean of the average of these 30 sets.

4.1 Cross-language article linking

The recall scores of the two datasets are 0.8953 and 0.8383, which is the upper bound of the system's performance. Since the candidate selection process relies heavily on translation, we think the difference between the two recall scores is due to the poor translation of unpopular content. We report the performance of ranking in terms of mean reciprocal rank (MRR).In Table 1, Column 2 (MRR) shows the performance of four feature configurations in the first dataset, BM25(Baseline), BM25+HT, BM25+ETO, and BM25+HT+ETO. To show CEEE's effectiveness, we list the performance before and after adding CEEE, including performance gain in parentheses. For the first dataset, we can see that CEEE boosts the baseline configuration by a significant margin. Using all three features, BM25+HT+ETO, the system achieves an MRR of 0.7967, which is the best score among all configurations. All configurations achieve statistically significant performance gains after adding CEEE.

Since the second dataset consists of seldom-read articles, we assume that some of the translations will be incorrect due to lack of popularity, and this may negatively impact system performance. Looking at the results of the second dataset in Table 2, we can see that HT and CEEE do not make the same improvement as they did in

the first dataset, since both of them heavily rely on Google Translate results. The MRR/recall ratio of the second dataset is slightly better than that of the first dataset with the BM25 configuration. We believe this is because there are more ambiguous articles in the first dataset than in the second dataset. For example, "The Hunger Games" is an article describing the trilogy of novels in Wikipedia. But its Baike candidates include several articles with identical titles, such as the first novel of the trilogy, the movies series, and video games. The results suggest that translation quality affects our system's performance, and that the candidate selection process is also impacted by the popularity of the query article.

In Table 3, we compare our work with the state-of-the-art system developed by (Wang et al., 2014). Our features improve performance relative to the baseline by 29.62% and 14.70% for the first and second datasets, respectively, which represents a statistically significant improvement over Wang et al.'s (2014) best system (26.86% and 12.81% relative improvement for the two datasets). It's worth noting that (Wang et al., 2014) utilized another feature based on topic models which requires known cross-language links.

All the experiments above treat the cosine similarity between the query and each candidate embedding as a feature for the SVM classifier. Next, we show the results on the first dataset when using the embedding vector as the feature vector for the SVM classifier. More specifically, the CEEE feature is a vector of 200 dimensions. We get an MRR of 0.6352 when the classifier only takes the embedding vector as input, which is a significant gain compared to the MRR of BM25. However, when CEEE is used to measure query-candidate cosine similarity, we only get an MRR of 0.4629. This suggests that CEEE has learned a sufficiently accurate mapping between Wikipedia and Baidu Baike, but there is still room for improvement.

5 Discussion

According to (Wang et al., 2014), one common CLAL error type is due to several articles having the same title. These same-title articles are entities belonging to different categories or one entity with several duplicate articles. Our best system configuration with the CEEE similarity feature fixed 21 such errors made by Wang et al.'s (2014) system. For example, for the English article Frankenstein (novel), Wang's system ranks the Chinese article 科学怪人(movie) as number one, but our system ranks the correct Chinese article 弗兰肯斯坦(novel) first. We then refer the set of these 21 errors as "Successfully corrected set (SCS)." Although our proposed CEEE feature can effectively disambiguate articles with the same title, it still failed to correct 90 of Wang's errors and generated 6 new ones. The set of the 90 uncorrected errors and the set of the six new errors are referred to as "Uncorrected set (US)" and "Mistakenly corrected set (MCS)," respectively. For example, the English article "British Museum" is still mistakenly linked to "英国科学博物馆(Science Museum, London)", when the correct corresponding article is in fact "大英博物馆".

We propose that the farther a concept is from the bilingual pairs used to train CEEE, the more likely it is to be linked to a non-corresponding article. To test this hypothesis, we calculated the link distance between each of the pairs above to the nearest training pair. Then, we calculate the average distance for SCS, US, and MCS. The results show that SCS has the least link distance, followed by US, and then MCS, which is consistent with our hypothesis. This finding suggests that in order to improve CLAL performance, we can introduce more cross-lingual pairs into the training data. In our future work, we plan to apply other lexical resources such as WordNet to find synonyms instead of simple string matching.

6 Conclusion

We describe a method to learn bilingual entity embedding for cross-encyclopedia CLAL. The embedding model is designed to encode both monolingual and bilingual entity structure so that related entities will be close to each other in the vector space. To acquire the training data without human annotation, we also offer a way to discover article correspondence among different encyclopedias. Our results show that using our proposed embedding outperforms the current best cross-encyclopedia CLAL system by statistically significant margin. Further investigations suggest that the proposed embedding can help to disambiguate candidates with the same title.

References

Nacéra Bennacer, Mia Johnson Vioulès, Maximiliano Ariel López, and Gianluca Quercini. 2015. *A*

Multilingual Approach to Discover Cross-Language Links in Wikipedia. Springer.

Zhiting Hu, Poyao Huang, Yuntian Deng, Yingkai Gao, and Eric P. Xing. 2015. Entity hierarchy embedding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1292–1300.

Taesung Lee and Seung-won Hwang. 2013. Bootstrapping entity translation on weakly comparable corpora. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 631–640. Association for Computational Linguistics.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Yuezhang Li, Ronghuo Zheng, Tian Tian, Zhiting Hu, Rahul Iyer, and Katia Sycara. 2016. Joint embedding of hierarchical categories and entities for concept categorization and dataless classification. *arXiv preprint arXiv:1607.07956*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 2013 Conference on Neural Information Processing Systems (NIPS)*, pages 3111–3119.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Dong Nguyen, Arnold Overwijk, Claudia Hauff, Dolf R. B. Trieschnigg, Djoerd Hiemstra, and Franciska de Jong. 2009. *WikiTranslate: Query Translation for Cross-Lingual Information Retrieval Using Only Wikipedia*.

Jong-Hoon Oh, Daisuke Kawahara, Kiyotaka Uchimoto, Jun'ichi Kazama, and Kentaro Torisawa. 2008. Enriching multilingual language resources by discovering missing cross-language links in wikipedia. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 322–328.

Philipp Sorg and Philipp Cimiano. 2008. Enriching the crosslingual link structure of wikipedia-a classification-based approach. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*, pages 49–54.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598.

Yu-Chun Wang, Chun-Kai Wu, and Richard Tzong-Han Tsai. 2014. Cross-language and cross-encyclopedia article linking using mixed-language topic model and hypernym translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 586–591. Association for Computational Linguistics.

Zhichun Wang, Juanzi Li, Zhigang Wang, and Jie Tang. 2012. Cross-lingual knowledge linking across wiki knowledge bases. In *Proceedings of the 21st international conference on World Wide Web*, pages 459–468.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. *arXiv preprint arXiv:1601.01343*.