Encyclolink: A Cross-Encyclopedia, Cross-language Article-Linking System and Web-based Search Interface

Yu-Chun Wang[1] **Ka Ming Wong**[2] **Chun-Kai Wu**[3] **Chao-Lin Pan**[2] **Richard Tzong-Han Tsai**[2*]

[1]Department of Buddhist Studies, Dharma Drum Institute of Liberal Arts, Taiwan

[2]Department of Computer Sciecne and Information Engineering, National Central University, Taiwan

[3]Department of Computer Sciecne, National Tsin Hua University, Taiwan

ycwang@dila.edu.tw marketforwkm@gmail.com

j3rmp4d93@gmail.com adhesivee@gmail.com thtsai@ncu.edu.tw

Abstract

Cross-language article linking (CLAL) is the task of finding corresponding article pairs across encyclopedias of different languages. In this paper, we present Encyclolink, a web-based CLAL search interface designed to help users find equivalent encyclopedia articles in Baidu Baike for a given English Wikipedia article title query. Encyclolink is powered by our cross-encyclopedia entity embedding CLAL system (0.8 MRR). The browser-based interface provides users with a clear and easily readable preview of the contents of retrieved articles for comparison.

1 Introduction

Online encyclopedias are among the most frequently used Internet services today, providing information summaries on millions of topics in all branches of knowledge. Wikipedia is one of the largest online encyclopedias and has many language versions, but there are alternatives to Wikipedia in some languages. In China, for example, Baidu Baike and Hudong are the largest encyclopedia sites. However, since the various encyclopedias in different languages have no connections to each other, it makes it difficult for searchers to find comprehensive results drawn from multiple online encyclopedias in multiple languages. If all or some of the world's online encyclopedias were integrated in a "meta-encyclopedia", potentially even translated into a user's local language, it would greatly richen search and information retrieval, helping users explore multiple viewpoints on a topic from users of other languages.

*corresponding author

In this paper, we introduce Encyclolink, a system that links articles written in two different languages from two different encyclopedia platforms, allowing users to search and browse information from two online encyclopedias in one interface. Encyclolink is based on cross-language article linking (CLAL), the task of creating links between equivalent articles written in different languages from different encyclopedias. Using Encyclolink, a user can input an English query and then see all linked Chinese articles in order of relevance. The user can then browse the English article and its Chinese counterpart side by side for comparison.

2 Related Work

Cross-language article linking is a new research target. The related work can be mainly divided into two groups: CLAL on Wikipedia and CLAL on different encyclopedias.

2.1 CLAL across Wikipedia Language Versions

The first work that aimed to find new cross-language links between English and German is Sorg and Cimiano (2008). They proposed a chain link hypothesis, which assumes that for any two cross-lingual linked articles, there are chain links in many language versions between them. They designed a candidate selection process based on the hypothesis and built a classifier with text-based features to predict the links.

Oh et al. (2008) later designed a language-independent approach. They first converted every English and Japanese Wikipedia article into vectors of the link, text and context of that article. Then they translated English vectors into Japanese using a dictionary created from existing cross-language links. Finally they adopted BM-25 to compute similarity between these vectors to select candidate links.

The Companion Volume of the IJCNLP 2017 Proceedings: System Demonstrations, pages 13–16,
Taipei, Taiwan, November 27 – December 1, 2017.

Wang et al. (2012), on the other hand, relied solely on link structure between English and Chinese Wikipedia articles. They found out that the more common links or categories there are between two cross-lingual articles, the more likely they are to be equivalent articles. They first created a graph for each Wikipedia version; nodes in the graph represent articles and edges are hyperlinks between articles. Then, they used the cross-language links between two Wikipedia versions to reconnect two graphs in a pair-wise connectivity graph (PCG), which served as the structure of their learning model.

2.2 CLAL between Wikipedia and Baidu Baike

Wang et al. (2012) attempted to integrate two different encyclopedias, English Wikipedia and Chinese Baidu Baike, into one cross-language encyclopedia. They created over 0.2 million links between the encyclopedias, but their approach requires many manually pre-linked article links and category links to create the pair-wise connectivity graph (PCG) model. Furthermore, in the paper, they do not mention how to verify accuracy of the newly discovered cross-language links.

Another relevant work is Wang et al. (2014), which also focuses on linking English Wikipedia and Baidu Baike articles. To select and predict article links, they designed text-related features for an SVM classifier. Their features include bidirectional title matching, title similarity, hypernym translation and English title occurrence.

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. 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 translation 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

Our model is based on Mikolov et al.'s (2013) skip-gram model. The training objective of the skip-gram model is to maximize the probability of predicting the target word given the context, where the target-context pairs are extracted by sliding a window over the entire corpus.

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. On the assumption that the entities mentioned in an article are somehow related to the article's meaning, for a given context article, we treat all entities mentioned in it as target entities. Given a set of target-context entity pairs $E = \{(t, c)\}$, we learn the embeddings of entities by maximizing the training objective:

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

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 an online encyclopedia:

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

, where $v_t, v_c \subset \mathbb{R}^d$ is the embedding of an entity, d is the embedding size and $\odot$ is dot product.

3.2 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. We train the model with (1) Baidu as target and Wikipedia as context, (2) Wikipedia as target and Baidu as context, (3) Wikipedia as both target and context, and (4) Baidu as both target and context. During task (3), only m^w is updated, and during task (4), only m^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.

3.3 Cross-Encyclopedia Entity Embedding Feature

After training, the learned embeddings 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. 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:

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

4 Demo System

4.1 System Architecture

Our web-based demo system, Encyclolink, is composed of two modules: Web UI and Web Service. The Web UI is mainly written in simple HTML with CSS. The Web UI takes a user's input as the query. After receiving a query, the UI sends it to the Web Service with a JavaScript function. Then the Web Service will call our main CLAL module, described in Section 3, to retrieve and rank the candidate articles according to their scores. The flowchart of Encyclolink is depicted in Figure 1.

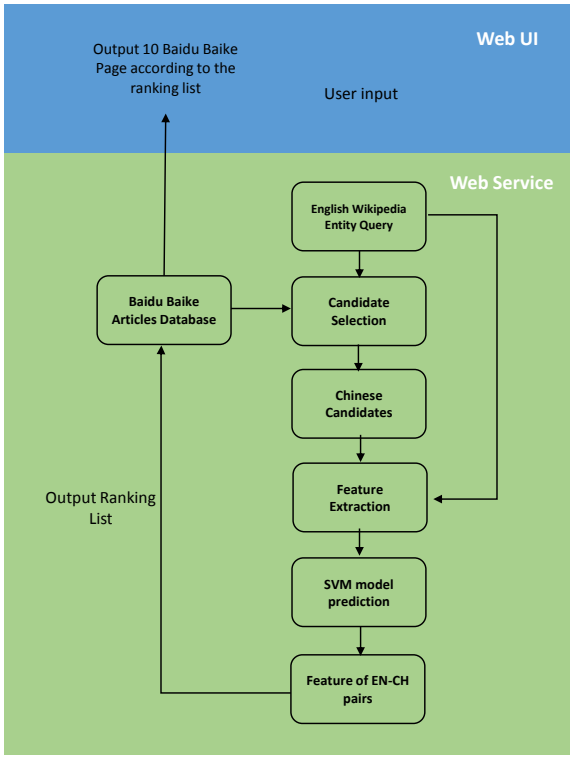

Figure 1: The flowchar of Encyclolink

4.2 User Interface and Application

The user interface of Encyclolink is separated into two main areas shown in Figure 2. The upper area contains a preview of the original English Wikipedia page. In the bottom area are the CLAL results given the input query. At the upper left corner of the UI screen, there is a text search field in which users can input a title of an article of interest. After the user presses the "Go" button, a preview of the English article will appear below the search field, including a hyperlink to the original English Wikipedia article. Meanwhile, the input query is sent to Web Service module, which selects and returns 10 candidate articles from Baidu Baike. They are listed in descending order according to their relevance scores to the query. Each of the candidates is also accompanied by a hyperlink, which users can click to open a window to view the contents of the Baidu Baike article. The UI allows users to view the English article alongside the corresponding Baidu Baike article on the same page for comparison.

5 Conclusion

This paper describes the Encyclolink system, a web-based system that can link articles from En-

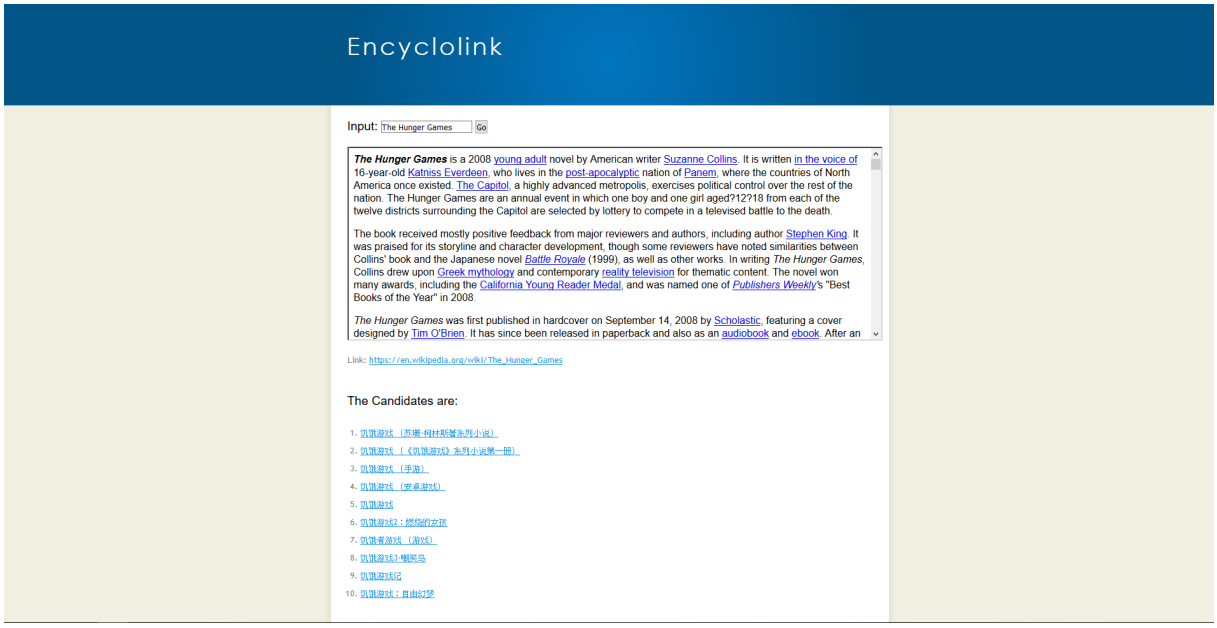

Figure 2: The Web UI of Encyclolink

glish Wikipedia and Chinese Baidu Baike, allowing users to retrieve and examine information from two of the world's largest online encyclopedias. A Encyclolink user can enter an English query to retrieve the matching English Wikipedia articles and a ranked list of corresponding Chinese articles in Baidu Baike. Encyclolink displays the results in a simple preview interface that lets users compare both English and Chinese encyclopedia articles side by side.

Acknowledgement

This research was supported in part by the Ministry of Science and Technology of Taiwan (MOST 104-2221-E-008-034-MY3), National Taiwan University (NTU-106R104045), Intel Corporation, and Delta Electronics, and Advantech.

References

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.

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.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 1661–1670.

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)*. Association for Computational Linguistics, pages 586–591.

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.