Cross-language and Cross-encyclopedia Article Linking Using Mixed-language Topic Model and Hypernym Translation

Yu-Chun Wang
Department of CSIE
National Taiwan University
Taipei, Taiwan
d97023@csie.ntu.edu.tw

Chun-Kai Wu
Department of CSIE
National Tsinghua University
Hsinchu, Taiwan
s102065512@m102.nthu.edu.tw

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

Abstract

Creating cross-language article links among different online encyclopedias is now an important task in the unification of multilingual knowledge bases. In this paper, we propose a cross-language article linking method using a mixed-language topic model and hypernym translation features based on an SVM model to link English Wikipedia and Chinese Baidu Baike, the most widely used Wiki-like encyclopedia in China. To evaluate our approach, we compile a data set from the top 500 Baidu Baike articles and their corresponding English Wiki articles. The evaluation results show that our approach achieves 80.95% in MRR and 87.46% in recall. Our method does not heavily depend on linguistic characteristics and can be easily extended to generate cross-language article links among different online encyclopedias in other languages.

1 Introduction

Online encyclopedias are among the most frequently used Internet services today. One of the largest and best known online encyclopedias is Wikipedia. Wikipedia has many language versions, and articles in one language contain hyperlinks to corresponding pages in other languages. However, the coverage of different language versions of Wikipedia is very inconsistent. Table 1 shows the statistics of inter-language link pages in the English and Chinese editions in February 2014. The total number of Chinese articles is about one-quarter of English ones, and only 2.3% of English articles have inter-language links to their Chinese versions.

	Articles	Inter-language Links		Ratio
zh	755,628	zh2en	486,086	64.3%
en	4,470,246	en2zh	106,729	2.3%

Table 1: Inter-Language Links in Wikipedia

However, there are alternatives to Wikipedia for some languages. In China, for example Baidu Baike and Hudong are the largest encyclopedia sites, containing more than 6.2 and 7 million Chinese articles respectively. Similarly, in Korea, Naver Knowledge Encyclopedia has a large presence.

Since alternative encyclopedias like Baidu Baike are larger (by article count) and growing faster than the Chinese Wikipedia, it is worthwhile to investigate creating cross-language links among different online encyclopedias. Several works have focused on creating cross-language links between Wikipedia language versions (Oh et al., 2008; Sorg and Cimiano, 2008) or finding a cross-language link for each entity mention in a Wikipedia article, namely Cross-Language Link Discovery (CLLD) (Tang et al., 2013; McNamee et al., 2011). These works were able to exploit the link structure and metadata common to all Wikipedia language versions. However, when linking between different online encyclopedia platforms this is more difficult as many of these structural features are different or not shared. To date, little research has been done into linking between encyclopedias on different platforms.

Title translation is an effective and widely used method of creating cross-language links between encyclopedia articles. (Wang et al., 2012; Adafre and de Rijke, 2005) However, title translation alone is not always sufficient. In some cases, for example, the titles of corresponding articles in different languages do not even match. Other methods must be used along with title translation to create a more robust linking tool.

[*] corresponding author

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 586–591,
Baltimore, Maryland, USA, June 23-25 2014.

In this paper, we propose a method comprising title and hypernym translation and mixed-language topic model methods to select and link related articles between the English Wikipedia and Baidu Baike online encyclopedias. We also compile a suitable dataset from the above two encyclopedias to evaluate the linking accuracy of our method.

2 Method

Cross-language article linking between different encyclopedias can be formulated as follows: For each encyclopedia K, a collection of human-written articles, can be defined as $K = \{a_i\}_{i=1}^{n}$, where a_i is an article in K and n is the size of K. Article linking can then be defined as follows: Given two encyclopedia K_1 and K_2, cross-language article linking is the task of finding the corresponding equivalent article a_j from encyclopedia K_2 for each article a_i from encyclopedia K_1. Equivalent articles are articles that describe the same topic in different languages.

Our approach to cross-language article linking comprises two stages: candidate selection, which produces a list of candidate articles, and candidate ranking, which ranks that list.

2.1 Candidate Selection

Since knowledge bases (KB) may contain millions of articles, comparison between all possible pairs in two knowledge bases is time-consuming and sometimes impractical. To avoid brute-force comparison, we first select plausible candidate articles on which to focus our efforts. To extract possible candidates, two similarity calculation methods are carried out: title matching and title similarity.

2.1.1 Title Matching

In our title matching method, we formulate candidate selection as an English-Chinese cross-language information retrieval (CLIR) problem (Schönhofen et al., 2008), in which every English article's title is treated as a query and all the articles in the Chinese encyclopedia are treated as the documents. We employ the two main CLIR methods: query translation and document translation.

In query translation, we translate the title of every English article into Chinese and then use these translated titles as queries to retrieve articles from the Chinese encyclopedia. In document translation, we translate the contents of the entire Chinese encyclopedia into English and then search them using the original English titles. The top 100 results for the query-translation and the top 100 results for document-translation steps are unionized. The resulting list contains our title-matching candidates.

For the query- and document-translation steps, we use the Lucene search engine with similarity scores calculated by the Okapi BM25 ranking function (Beaulieu et al., 1997). We separate all words in the translated and original English article titles with the "OR" operator before submission to the search engine. For all E-C and C-E translation tasks, we use Google Translate.

2.1.2 Title Similarity

In the title similarity method, every Chinese article title is represented as a vector, and each distinct character in all these titles is a dimension of all vectors. The title of each English article is translated into Chinese and represented as a vector. Then, cosine similarity between this vector and the vector of each Chinese title is measured as title similarity.

2.2 Candidate Ranking

The second stage of our approach is to score each viable candidate using a supervised learning method, and then sort all candidates in order of score from high to low as final output.

Each article x_i in KB K_1 can be represented by a feature vector $\mathbf{x}_i = (f_1(x_i), f_2(x_i), \ldots, f_n(x_i))$. Also, we have $\mathbf{y}_j = (f_1(y_j), f_2(y_j), \ldots, f_n(y_j))$ for a candidate article y_j in KB K_2. Then, individual feature functions $F_k(x_i, y_j)$ are based on the feature properties of both article a_i and a_j. The top predicted corresponding article y_j in the knowledge base K_2 for an input article x_i in K_1 should receive a higher score than any other entity in $K_2, a_m \in K_2, m \neq j$. We use the support vector machine (SVM) approach to determine the probability of each pair $(\mathbf{x}_i, \mathbf{y}_j)$ being equivalent. Our SVM model's features are described below.

Title Matching and Title Similarity Feature (Baseline)

We use the results of title matching and title similarity from the candidate selection stage as two features for the candidate ranking stage. The similarity values generated by title matching and title similarity are used directly as real value features in the SVM model.

Mixed-language Topic Model Feature (MTM)

For a linked English-Chinese article pair, the distribution of words used in each usually shows some convergence. The two semantically corresponding articles often have many related terms, which results in clusters of specific words. If two articles do not describe the same topic, the distribution of terms is often scattered. (Misra et al., 2008) Thus, the distribution of terms is good measurement of article similarity.

Because the number of all possible words is too large, we adopt a topic model to gather the words into some latent topics. For this feature, we use the Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA can be seen as a typical probabilistic approach to latent topic computation. Each topic is represented by a distribution of words, and each word has a probability score used to measure its contribution to the topic. To train the LDA model, the pair English and Chinese articles are concatenated into a single document. English and Chinese terms are all regarded as terms of the same language and the LDA topic model, namely mixed-language topic model, generates both English and Chinese terms for each latent topic. Then, for each English article and Chinese candidate pair in testing, the LDA model provides the distribution of the latent topics. Next, we can use entropy to measure the distribution of topics. The entropy of the estimated topic distribution of a related article is expected to be lower than that of an unrelated article. We can calculate the entropy of the distribution as a value for SVM. The entropy is defined as follows:

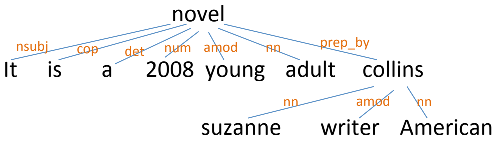
$$H = -\sum_{j=1}^{T} \vec{\theta_{dj}} \log \vec{\theta_{dj}}$$

where T is the number of latent topics, θ_{dj} is the topic distribution of a given topic j.

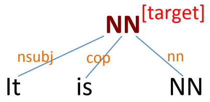
Hypernym Translation Feature (HT)

The first sentence of an encyclopedia article usually contains the title of the article. It may also contain a hypernym that defines the category of the article. For example, the first sentence of the "iPad" article in the English Wikipedia begins, "iPad is a line of tablet computers designed and marketed by Apple Inc..." In this sentence, the term "tablet computers" is the hypernym of iPad. These extracted hypernyms can be treated as article categories. Therefore, articles containing the same hypernym are likely to belong to the same category.

In this study, we only carry out title hypernym extraction on the first sentences of English articles due to the looser syntactic structure of Chinese. To generate dependency parse trees for the sentences, we adopt the Stanford Dependency Parser. Then, we manually designed seven patterns to extract hypernyms from the parse tree structures. To demonstrate this idea, let us take the English article "The Hunger Games" for example. The first sentence of this article is "The Hunger Games is a 2008 young adult novel by American writer Suzanne Collins." Since article titles may be named entities or compound nouns, the dependency parser may mislabel them and thus output an incorrect parse tree. To avoid this problem, we first replace all instances of an article's title in the first sentence with pronouns. For example, the previous sentence is rewritten as "It is a 2008 young adult novel by American writer Suzanne Collins." Then, the dependency parser generates the following parse tree:



Next, we apply our predefined syntactic patterns to extract the hypernym. (Hearst, 1992) If any pattern matches the structure of the dependency parse tree, the hypernym can be extracted. In the above example, the following pattern is matched:



In this pattern, the rightmost leaf is the hypernym target. Thus, we can extract the hypernym "novel" from the previous example. The term "novel" is the extracted hypernym of the English article "The Hunger Games".

After extracting the hypernym of the English article, the hypernym is translated into Chinese. The value of this feature in the SVM model is calculated as follows:

$$F_{hypernym}(h) = \log count(translated(h))$$

where h is the hypernym, $translated(h)$ is the Chinese translation of the term h.

English Title Occurrence Feature (ETO)

In a Baidu Baike article, the first sentence may contain a parenthetical translation of the main title. For example, the first sentence of the Chinese article on San Francisco is “旧金山（San Francisco），又译‘圣弗朗西斯科’、‘三藩市’。”. We regard the appearance of the English title in the first sentence of a Baidu Baike article as a binary feature: If the English title appears in the first sentence, the value of this feature is 1; otherwise, the value is 0.

3 Evalutaion

3.1 Evaluation Dataset

In order to evaluate the performance of cross-language article linking between English Wikiepdia and Chinese Baidu Baike, we compile an English-Chinese evaluation dataset from Wikipedia and Baidu Baike online encyclopedias. First, our spider crawls the entire contents of English Wikipedia and Chinese Baidu Baike. Since the two encyclopedias' article formats differ, we copy the information in each article (title, content, category, etc.) into a standardized XML structure. In order 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.

To evaluate the performance of our method on linking different types of encyclopedia articles, we compile a set containing the most popular articles. 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.

Because our approach uses an SVM model, the data set should be split into training and test sets. For statistical generality, each 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 evaluation sets.

3.2 Evaluation Metrics

To measure the quality of cross-language entity linking, we use the following three metrics. For each English article queries, ten output Baidu Baike candidates are generated in a ranked list. To define the metrics, we use following notations: N is the number of English query; $r_{i,j}$ is j-th correct Chinese article for i-th English query; $c_{i,k}$ is k-th candiate the system output for i-th English query.

Top-k Accuracy (ACC)

ACC measures the correctness of the first candidate in the candidate list. $ACC = 1$ means that all top candidates are correctly linked (i.e. they match one of the references), and $ACC = 0$ means that none of the top candidates is correct.

$$ACC = \frac{1}{N}\sum_{i=1}^{N}\left\{\begin{array}{ll} 1 & \text{if } \exists r_{i,j} : r_{i,j} = c_{i,k} \\ 0 & \text{otherwise} \end{array}\right\}$$

Mean Reciprocal Rank (MRR)

Traditional MRR measures any correct answer produced by the system from among the candidates. 1/MRR approximates the average rank of the correct transliteration. An MRR closer to 1 implies that the correct answer usually appears close to the top of the n-best lists.

$$\begin{aligned} RR_i &= \left\{\begin{array}{ll} \min_j \frac{1}{j} & \text{if } \exists r_{i,j}, c_{i,k} : r_{i,j} = c_{i,k} \\ 0 & \text{otherwise} \end{array}\right\} \\ MRR &= \frac{1}{N}\sum_{i=1}^{N} RR_i \end{aligned}$$

Recall

Recall is the fraction of the retrieved articles that are relevant to the given query. Recall is used to measure the performance of the candidate selection method. If the candidate selection method can actually select the correct Chinese candidate, the recall will be high.

$$Recall = \frac{|\text{relevant articles}| \cap |\text{retrieved articles}|}{|\text{relevant articles}|}$$

3.3 Evaluation Results

The overall results of our method achieves 80.95% in MRR and 87.46% in recall. Figure 1 shows the top-k ACC from the top 1 to 5. These results show that our method is very effective in linking articles in English Wikipedia to those in Baidu Baike.

In order to show the benefits of each feature used in the SVM model, we conduct a experiment to test the performance of different feature combinations. Because title similarity of the articles is a widely used method, we choose English and Chinese title similarity as the baseline. Then, another feature is added to each configuration until all the features have been added. Table 2 shows the final results of different feature combinations.

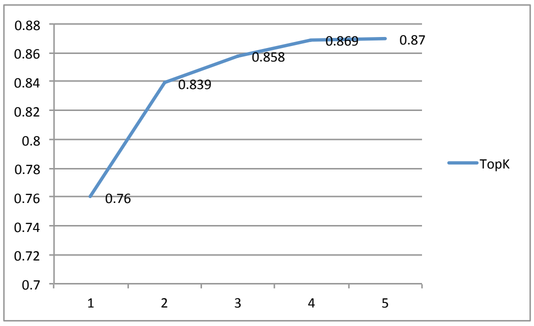


Figure 1: Top-k Accuracy

Level	Configuration	MRR
0	Baseline (BL)	0.6559
1	BL + MTM^{*1}	0.6967†
	BL + HT^{*2}	0.6975†
	BL + ETO^{*3}	0.6981†
2	BL + MTM + HT	0.7703†
	BL + MTM + ETO	0.7558†
	BL + HT + ETO	0.7682†
3	BL + MTM + HT + ETO	**0.8095**†

^{*1}MTM: mix-language topic model

^{*2}HT: hypernym translation

^{*3}ETO: English title occurrence

† This config. outperforms the best config. in last level with statistically significant difference.

Table 2: MRRs of Feature Combinations

In the results, we can observe that mix-language topic model, hypernym, and English title occurence features all noticeably improve the performance. Combining two of these three feature has more improvement and the combination of all the features achieves the best.

4 Discussion

Although our method can effectively generate cross-language links with high accuracy, some correct candidates are not ranked number one. After examining the results, we can divide errors into several categories:

The first kind of error is due to large literal differences between the English and Chinese titles. For example, for the English article “Nero”, our approach ranks the Chinese candidate “尼禄王” (“King Nero”) as number one, instead of the correct answer “尼禄·克劳狄乌斯·德鲁苏斯·日耳曼尼库斯” (the number two candidate). The title of the correct Chinese article is the full name of the Roman Emperor Nero (Nero Claudius Drusus Germanicus). The false positive “尼禄王” is a historical novel about the life of the Emperor Nero. Because of the large difference in title lengths, the value of the title similarity feature between the English article “Nero” and the corresponding Chinese article is low. Such length differences may cause the SVM model to rank the correct answer lower when the difference of other features are not so significant because the contents of the Chinese candidates are similar.

The second error type is caused by articles that have duplicates in Baidu Baike. For example, for the English article “Jensen Ackles”, our approach generates a link to the Chinese article “Jensen” in Baidu Baike. However, there is another Baidu article “詹森·阿克斯” (“Jensen Ackles”). These two articles both describe the actor Jensen Ackles. In this case, our approach still generates a correct link, although it is not the one in the gold standard.

The third error type is translation errors. For example, the English article “Raccoon” is linked to the Baidu article “狸” (raccoon dog), though the correct one is “浣熊” (raccoon). The reason is that Google Translate provides the translation “狸” instead of “浣熊”.

5 Conclusion

Cross-language article linking is the task of creating links between online encyclopedia articles in different languages that describe the same content. We propose a method based on article hypernym and topic model to link English Wikipedia articles to corresponding Chinese Baidu Baike articles. Our method comprises two stages: candidate selection and candidate ranking. We formulate candidate selection as a cross-language information retrieval task based on the title similarity between English and Chinese articles. In candidate ranking, we employ several features of the articles in our SVM model. To evaluate our method, we compile a dataset from English Wikipedia and Baidu Baike, containing the 500 most popular Baidu articles. Evaluation results of our method show an MRR of up to 80.95% and a recall of 87.46%. This shows that our method is effective in generating cross-language links between English Wikipedia and Baidu Baike with high accuracy. Our method does not heavily depend on linguistic characteristics and can be easily extended to generate cross-language article links among different encyclopedias in other languages.

References

Sisay Fissaha Adafre and Maarten de Rijke. 2005. Discovering missing links in wikipedia. In *Proceedings of the 3rd international workshop on Link discovery (LinkKDD '05)*.

M. Beaulieu, M. Gatford, X. Huang, S. Robertson, S. Walker, and P. Williams. 1997. Okapi at TREC-5. In *Proceedings of the fifth Text REtrieval Conference (TREC-5)*, pages 143–166.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, volume 2.

Paul McNamee, James Mayfield, Dawn Lawrie, Douglas W Oard, and David S Doermann. 2011. Cross-language entity linking. In *Proceedings of International Joint Con-ference on Natural Language Processing (IJCNLP)*, pages 255–263.

Hemant Misra, Olivier Cappe, and François Yvon. 2008. Using lda to detect semantically incoherent documents. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL '08)*.

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.

Pèter Schönhofen, Andràs Benczùr, Istvàn Bìrò, and Kàroly Csalogàny. 2008. Cross-language retrieval with wikipedia. *Advances in Multilingual and Multimodal Information Retrieval, Lecture Notes in Computer Science*, 5152:72–79.

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.

Ling-Xiang Tang, In-Su Kang, Fuminori Kimura, Yi-Hsun Lee, Andrew Trotman, Shlomo Geva, and Yue Xu. 2013. Overview of the ntcir-10 cross-lingual link discovery task. In *Proceedings of the Tenth NTCIR Workshop Meeting*.

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 (WWW '12)*.