Cross-lingual Link Discovery between Chinese and English Wiki Knowledge Bases

Qingliang Miao, Huayu Lu, Shu Zhang, Yao Meng
Fujitsu R&D Center Co., Ltd.
No.56 Dong Si Huan Zhong Rd, Chaoyang District, Beijing P.R. China
{qingliang.miao, zhangshu, mengyao}@cn.fujitsu.com
lvhuayu@gmail.com

Abstract

Wikipedia is an online multilingual encyclopedia that contains a very large number of articles covering most written languages. However, one critical issue for Wikipedia is that the pages in different languages are rarely linked except for the cross-lingual link between pages about the same subject. This could pose serious difficulties to humans and machines who try to seek information from different lingual sources. In order to address above issue, we propose a hybrid approach that exploits anchor strength, topic relevance and entity knowledge graph to automatically discovery cross-lingual links. In addition, we develop CELD, a system for automatically linking key terms in Chinese documents with English Concepts. As demonstrated in the experiment evaluation, the proposed model outperforms several baselines on the NTCIR data set, which has been designed especially for the cross-lingual link discovery evaluation.

1 Introduction

Wikipedia is the largest multilingual encyclopedia online with over 19 million articles in 218 written languages. However, the anchored links in Wikipedia articles are mainly created within the same language. Consequently, knowledge sharing and discovery could be impeded by the absence of links between different languages. Figure 1 shows the statistics of monolingual and cross-lingual alignment in Chinese and English Wikipedia. As it can be seen that there are 2.6 millions internal links within English Wikipedia and 0.32 millions internal links within Chinese Wikipedia, but only 0.18 millions links between Chinese Wikipedia pages to English ones. For example, in Chinese Wikipedia page "武术(Martial arts)", anchors are only linked to related Chinese articles about different kinds of martial arts such as "拳击(Boxing)", "柔道(Judo)" and "击剑(Fencing)". But, there is no anchors linked to other related English articles such as "Boxing", "Judo and "Fencing". This makes information flow and knowledge propagation could be easily blocked between articles of different languages.

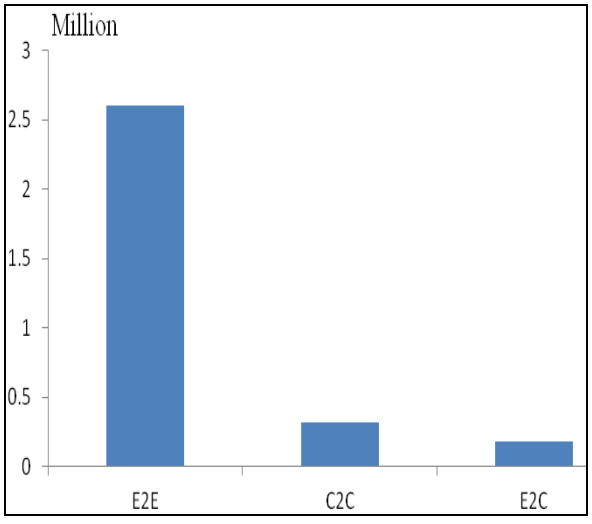

Figure 1. Statistics of English to English links (E2E), Chinese to Chinese links (C2C) and Chinese to English links (C2E).

Consequently, automatically creating cross-lingual links between Chinese and English Wikipedia would be very useful in information flow and knowledge sharing. At present, there

27th Pacific Asia Conference on Language, Information, and Computation pages 374－381

are several monolingual link discovery tools for English Wikipedia, which assist topic curators in discovering prospective anchors and targets for a given Wikipedia pages. However, no such cross-lingual tools yet exist, that support the cross-lingual linking of documents from multiple languages (Tang et al., 2012). As a result, the work is mainly taken by manual, which is obviously tedious, time consuming, and error prone.

One way to solve above issue is cross-lingual link discovery technology, which automatically creates potential links between documents in different languages. Cross-lingual link discovery not only accelerates the knowledge sharing in different languages on the Web, but also benefits many practical applications such as information retrieval and machine translation (Wang et al., 2012). In existing literature, a few approaches have been proposed for linking English Wikipedia to other languages (Kim and Gurevych, 2011; Fahrni et al., 2011). Generally speaking, there are three steps for Cross-lingual link discovery: (1) Apply information extraction techniques to extract key terms from source language documents. (2) Utilize machine translation systems to translate key terms and source documents into target language. (3) Apply entity resolution methods to identify the corresponding concepts in target language. However, in key term extraction step, most works rely on statistical characteristics of anchor text (Tang et al., 2012), but ignore the topic relevance. In this case, common concepts are selected as key terms, but these terms are not related to the topic of the Wikipedia page. For example, in Chinese Wikipedia page "武术 (Martial arts)", some countries' name such as "中国 (China)", "日本 (Japan)" and "韩国 (Korea)" are also selected as key terms when using anchor statistics. For term translation, existing methods usually depends on machine translation, and suffers from translation errors, particularly those involving named entities, such as person names (Cassidy et al., 2012). Moreover, machine translation systems are prone to introduce translation ambiguities. In entity resolution step, some works use simple title matching to find concept in target languages, which could not distinguish ambiguous entities effectively (Kim and Gurevych, 2011).

In this paper, we try to investigate the problem of cross-lingual link discovery from Chinese Wikipedia pages to English ones. The problem is non-trivial and poses a set of challenges.

Linguistic complexity

Chinese Wikipedia is more complex, because contributors of Chinese Wikipedia are from different Chinese spoken geographic areas and language variations. For example, Yue dialect[1] is a primary branch of Chinese spoken in southern China and Wu [2] is a Sino-Tibetan language spoken in most of southeast. Moreover, these contributors cite modern and ancient sources combining simplified and traditional Chinese text, as well as regional variants (Tang et al., 2012). Consequently, it is necessary to normalize words into simple Chinese before cross-lingual linking.

Key Term Extraction

There are different kinds of key term ranking methods that could be used in key term extraction, such as tf-idf, information gain, anchor probability and anchor strength (Kim and Gurevych, 2011). How to define a model to incorporate both the global statistical characteristics and topically related context together?

Translation

Key term translation could rely on bilingual dictionary and machine translation. This kind of methods could obtain high precision, while suffer from low recall. When using larger dictionaries or corpus for translation, it is prone to introduce translation ambiguities. How to increase recall without introducing additional ambiguities?

In order to solve the above challenges, we investigate several important factors of cross-lingual link discovery problem and propose a hybrid approach to solve the above issues. Our contributions include:

(1) We develop a normalization lexicon for Chinese variant character. This lexicon could be used for traditional and simplified Chinese transformation and other variations normalization. We also discovery entity knowledge from Wikipedia, Chinese encyclopedia, and then we build a knowledge graph that includes mentions, concepts, translations and corresponding confidence scores.

(2) We present an integrated model for key terms extraction, which leverages anchor

[1] http://zh-yue.wikipedia.org
[2] http://wuu.wikipedia.org

statistical probability information and topical relevance. Efficient candidate selection method and distinguishing algorithm enable this model meet the real-time requirements.

(3) We implement a system and evaluate it using NTCIR cross-lingual links discovery dataset. Comparing with several baselines, our system achieves high precision and recall.

The remainder of the paper is organized as follows. In the following section we review the existing literature. Then, we formally introduce the problem of cross-lingual link discovery and some related concepts in section 3. We introduce the proposed approach in section 4. We conduct comparative experiments and present the experiment results in section 5. At last, we conclude the paper with a summary of our work and give our future working directions.

2 Related Works

Generally speaking, link discovery is a kind of semantic annotation (Kiryakov et al., 2004), which is characterized as the dynamic creation of interrelationships between concepts in knowledge base and mentions in unstructured or semi-structured documents (Bontcheva and Rout, 2012).

In particular, most existing monolingual semantic annotation (MLSA) approaches annotate documents with links to Wikipedia or DBPedia. Mihalcea and Csomai (2007) first attempt to use Wikipedia to annotate monolingual text is their Wikify system. Wikify system includes two main steps, key term detection and disambiguation. The system identifies key terms according to link probabilities obtained from Wikipedia pages. In order to link key term to the appropriate concept, Wikify extracts features from the key term and its context, and compares these features to training examples obtained from the Wikipedia. Milne and Witten (2008) implement a similar system called Wikipedia Miner, which adopts supervised disambiguation approach using Wikipedia hyperlinks as training data. There are also some semantic annotation contests. For example, TAC's entity linking task[3] focuses on the linkage of named entities such as persons, organizations and geo-political entities to English Wikipedia concepts. Given a query that consists of a name string and a background document ID, the system is required to provide the ID of the knowledge base entry to which the name refers; or NIL if there is no such knowledge base entry. Due to the intrinsic ambiguity of named entities, most works in entity linking task focus on named entity disambiguation. For example, Han and Sun (2012) propose a generative entity-topic model that effectively joins context compatibility and topic coherence. Their model can accurately disambiguate most mentions in a document using both the local information and the global consistency.

Following this research stream, researchers have been paying more and more attention on cross-lingual semantic annotation (CLSA). Knowledge Base Population (KBP2011) evaluations propose a cross-lingual entity link task, which aims to find link between Chinese queries and English concepts. NTCIR9 cross-lingual link discovery task is another kind of cross-lingual semantic annotation. These two tasks are different in query selection criteria, leading to different technical difficulties and concerns. In KBP2011, key terms are manually selected to cover many ambiguous entities and name variants. Consequently, disambiguation is crucial in KBP2011. While in NTCIR9, participants have to extract key terms from given documents first. Since these extracted key terms are less ambiguous than KBP's entities, disambiguation has less effect on final performance (Kim and Gurevych, 2011). In contrast, translation plays an important role in NTCIR9 task. Another direction is cross-lingual knowledge linking across web knowledge bases. Wang et al. (2012) study the problem of creating cross-lingual links between English Wikipedia and Chinese encyclopedia Baidu Baike[4] and propose a linkage factor graph model.

Although CLSA is a new task, efforts in MLSA could be adopted. In particular, there are two conventional way to extend MLSA systems to the cross-lingual setting: the first one is applying MLSA method to link source language entity mentions to source language knowledge base concepts, and then link the source language knowledge base concepts to the corresponding target language knowledge base concepts. This strategy relies heavily on the existence of a reliable mapping between source language knowledge base and target language knowledge base. The second one is utilizing machine translation techniques to translate the source

[3] http://www.nist.gov/tac/2012/KBP/workshop/index.html

[4] http://baike.baidu.com/

language document or mentions into the target language, and then apply a MLSA method in the target language side. This process relies on machine translation output, and it will suffer from translation errors inevitably, particularly those involving named entities (Cassidy et al., 2012). In this paper, we leverage anchor probability information and topic relevance to extract key terms from Chinese documents. And then, we build a knowledge graph, and use this graph to translate key terms to English. Finally, cross-lingual links are identified by concept resolution model.

3 Problem Definition

In this section, we define the problem of cross-lingual link discovery and some related concepts.

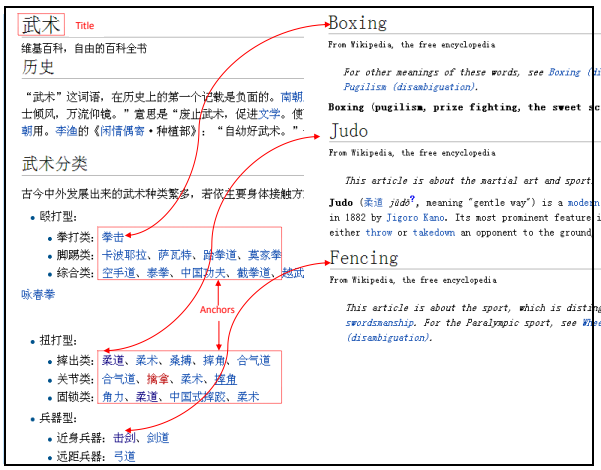

Figure 2. An Example of cross-lingual link discovery, Chinese to English links (C2E).

Definition 1: Wikipedia Knowledge Base

Wikipedia knowledge base is a collection of collaboratively written articles, each of which defines a specific concept. It can be formally represented as $K=\{a_i\}, i \in [1,n]$, where a_i is an article in K and n is the size of K. Each article a_i describes a specific concept. Each article includes four key elements, title name, textual content, anchors and categories and can be represented as $\{N(a_i), T(a_i), A(a_i), C(a_i)\}$, where $N(a_i)$ and $T(a_i)$ are the title name and textual content of the article a_i respectively; $A(a_i)$ denotes the set of anchors of the a_i, and $C(a_i)$ is the category tags of a_i.

Definition 2: Topic document

The topic documents are actual Wikipedia articles selected for link discovery. Anchors in topic documents are removed in the test data. For example, in Chinese to English link discovery task. Topic documents are Chinese articles without existing anchors. Topic document could be represented as $\{N(a_i), T(a_i), C(a_i)\}$, where $N(a_i)$ is the title name of the document, $T(a_i)$ is the textual content of the document a_i; and $C(a_i)$ is the category tags of the document a_i.

Definition 3: Anchor

An anchor is a piece of text that is relevant to the topic and worthy of being linked to other articles for further reading. Anchor text usually gives the user relevant descriptive or contextual information about the content of the link's destination.

Definition 4: Cross-lingual Link

Given one topic t in source language and a Wikipedia knowledge base K in target language, cross-lingual link discovery is the process of finding potential anchors in t and link to appropriate articles in K.

As shown in Figure 2, the topic "武术(Martial arts)" is from Chinese Wikipedia documents. There is no anchors (cross-lingual link) from topic "武术" to English Wikipedia articles. In the cross-lingual link discovery problem, our goal is to extract anchors such as "拳击(Boxing)", "柔道(Judo)" and "击剑(Fencing)", and then find semantic equivalent articles for all the extracted anchors in English Wikipedia knowledge base.

4 The Approach

In this section, we will first introduce the overview of the system. And then, we present key term extraction and translation and concept resolution.

4.1 System Overview

Figure 3 illustrates the overview of the cross-lingual link discovery system. The inputs of the system are Chinese topic documents and English Wikipedia knowledge base, and the outputs are anchors of Chinese topic documents and their linking concepts in English Wikipedia knowledge base.

The system consists of four parts: (1) key term extraction module (KEM); (2) knowledge mining module (KMM); (3) key term translation module (KTM) and (4) concept resolution module (CRM).

KEM first extracts key term candidates from the main text of Chinese topic documents. And

then, KEM refines key term candidates according to anchor statistical probability and topic similarity. Finally, key terms are normalized by normalization lexicon.

KMM extracts mentions, concepts and translations from Wikipedia dumps and Chinese encyclopedia. Then translation of concept is obtained by cross-lingual links and heuristic patterns. Finally, KMM builds knowledge graph including mentions, concepts and translations with corresponding confidence.

KTM has two inputs, one is key terms from KEM and the other one is knowledge graph from KMM. KTM first map key term (mention) to corresponding concept, and then find the translation of concept. In case we cannot find the mentions in the knowledge graph, we use machine translation systems to translate the key terms.

CRM first searches concept candidates from knowledge graph. This process could also be viewed as query expansion. After that, CRM ranks the concept candidates according to weighted sum of similarities including lexical similarity, local context similarity and category similarity. And then, CRM selects the one with highest similarity score as the final linking target and generates cross-lingual links.

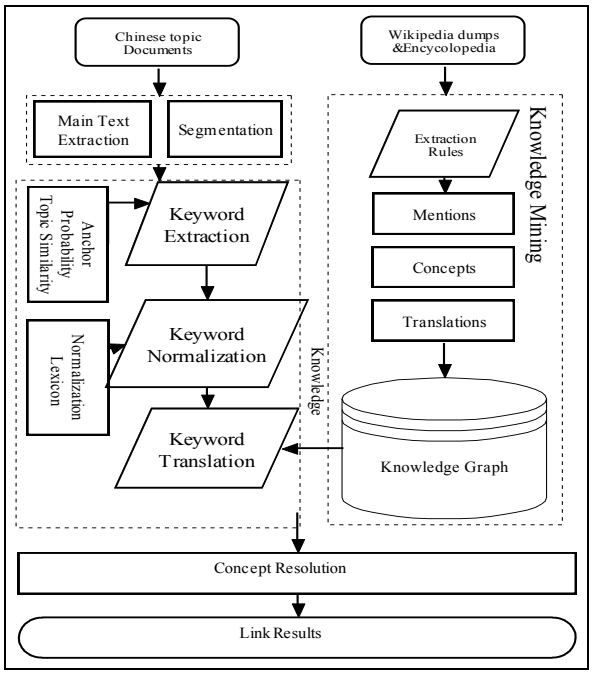

Figure 3. Overview of the cross-lingual link discovery system.

4.2 Key Term Extraction

In this section, we introduce the method for key term extraction. Key term extraction includes three steps: (1) key term candidate extraction from Chinese topic document; (2) key term candidate ranking according to importance and topic relevance; (3) key term normalization.

Kim and Gurevych (2011) introduce several key term candidate selection methods, such as noun phrases, named entities and anchor text. They also present some key term candidate ranking method such as tf-idf, anchor probability. In order to obtain topic-related and important terms, we leverage anchor strength and topic relatedness to rank key term candidates in this paper. In particular, we extract all n-grams of size 1 to 5, because n-grams subsume most key term candidates, which could obtain a high recall. Then, we compute anchor strength and topic relevance. Anchor strength measures the probability of the given text being used as an anchor text to its most frequent target in the Wikipedia corpus. Anchor strength could be computed as follows:

$$anchorStrength=\frac{count(c,d_{anchor})}{count(c,d)} \quad (1)$$

where $count(c.d_{anchor})$ denotes the count of anchor candidate c being used as an anchor in a document d, and $count(c.d)$ is the count of c appearing in a document d. In this paper, we filter out the key term candidates whose anchor strength is low than 0.001.

Topic relevance is computed as follows:

$$relatedness(a,b)=\frac{\log(\max(|A|,|B|))-\log(|A|\cap|B|)}{\log(|W|)-\log(\min(|A|,|B|))} \quad (2)$$

where a and b are two articles, A and B are the sets of all articles that link to a and b respectively, and W is set of all articles in Wikipedia. In this paper, we compute the semantic relatedness between each key term candidate and the topic. In particular, we first map the key term candidate to its corresponding concept, and then compute the semantic relatedness with the topic. If the key term candidate does not have any associated concept, we discard it. For example, given the topic document t and key term candidate a, we first find the concept c of a, and then compute the semantic relatedness between t and c. Finally, we filter out some key term candidates whose semantic relatedness is low than a threshold.

After that, we normalize the key terms according to the normalization lexicon. The normalization lexicon is derived from Wiktionary[5], which is a multilingual, web-based project to create a free content dictionary,

[5] http://zh.wiktionary.org/zh/

available in 158 languages. The lexicon contains 4747 traditional and simple Chinese character pairs. Most key terms could be normalized by simply looking up the normalization lexicon except for some cases. For example, in phrase "乾燥 (Drying)", character "乾" should be convert to "干", while in phrase "乾隆 (Qianlong_Emperor)" character "乾" should not be convert to "干". For these special cases, we have to build another dictionary, which includes the special phrases.

4.3 Key Term Translation

In this section, we first introduce how to mine entity knowledge from Wikipedia dumps and Chinese encyclopedia. And then, we introduce the structure of the knowledge graph. Finally, we illustrate how to use this knowledge graph to translate key terms. In particular, the knowledge can be built in two steps:

(1) Extracting mentions and concepts;
(2) Extract concepts and corresponding translations.

KMM extracts mentions and corresponding concepts by using redirection links, anchor links and pre-defined extraction patterns. Redirections in Wikipedia and encyclopedia are good indicators for mentions and concepts. Anchor links could also be used to trace to which concepts the mention links. In this paper, we use anchor links in Chinese Wikipedia and encyclopedia such as Baidu Baike and Hudong Baike[6]. We also exploit synonyms and linguistic patterns such as such as "A also called B", "A known as B", "A is referred to as B". After mention and concept extraction, we compute the confidence score that measures how confident the mention referring to concepts. For redirection links and linguist patterns, the confident score is assigned 1.0, since they are manually annotated. For anchor links, we assign the linking frequency as confident scores for corresponding mention and concept pairs.

KMM extracts concepts and their translations according to cross-lingual links and linguistic patterns. Cross-lingual links connect articles on the same concept in different languages, therefore concept and their translation pairs could be extracted. Besides cross-lingual links, we also discovery translations from Chinese encyclopedia through linguistic patterns, such as "A's English name is B", "A's abbreviation is C". The confident scores are set to 1.0.

After that, we built mention, concept and translation graph MCTG. MCTG includes mention, concept and translation layers. The associations between different layers are represented as interlayer links, and each association is assigned a confident score.

In key term translation, we adopt a cascade translation strategy. For a key term (mention), we first obtain the corresponding concepts and their confident scores. Then, we search the graph to find the translations for each concept. If the knowledge graph does not contain the mention, concept or translation, we use a machine translation system to translate the mention. Figure 4 illustrates a translation example. Given a mention such as "和阗玉" or "昆仑玉", we first find corresponding concept "和田玉", and then map the concept "和田玉" to its translation "Hetian jade" and "nephrite".

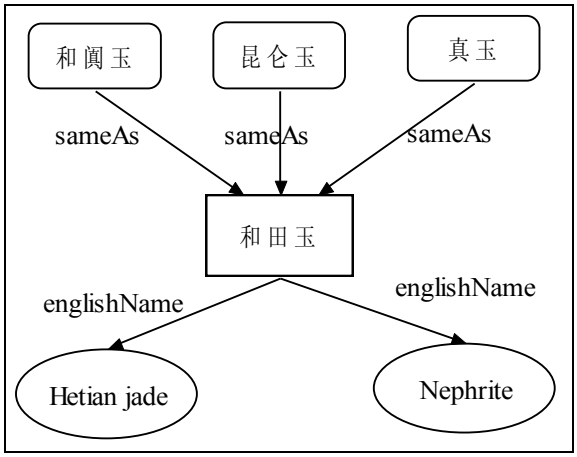

Figure 4. A key term translation example

4.4 Concept Resolution

After key term translation, we use the knowledge graph to select concept candidates for each mention and obtain a concept candidate set *S*. To identify the exact concept the key term refers, our system uses the weighted sum of similarities including lexical similarity, local context similarity and category similarity to determine which concept is the right one. In particular, we adopt Levenshtein distance[7] based algorithm to compute lexical similarity between mentions and concepts' titles. We also adopt vector-space model using bag-of-words to compute the textual similarity. Besides local similarity, we also

[6] http://www.baike.com/

[7] http://en.wikipedia.org/wiki/Levenshtein_distance

consider category similarity, for each concept candidate c_c in S, we find the English concept c_e whose title exactly matches the concept candidate. When multiple English concepts match the concept candidate, we find the most specific common class that subsumes c_c and c_e in the class taxonomy of Wikipedia. And then, we compute the path length between c_c and c_e . Finally, we select the one with largest similarity as the final linking target and generate cross-lingual links. In this work, the weight of each similarity is estimated from a manually collected training data set.

5 Experiments

In this section, we report a primary experiment aimed at evaluating the proposed method and system. We first describe the datasets used in our study and then we give experiment setup and results to demonstrate the effectiveness of our method for cross-lingual link discovery task.

5.1 Experimental Setup

In this experiment, we use the same dataset in (Tang et al., 2012), which is provided by NTCIR. The dumps of the Chinese and English Wikipedia are downloaded in June 2010. There are 3,484,250 English articles and 316,251 Chinese articles respectively. The test data contains a set of 36 Chinese topics[8]. The ground-truth is derived from Wikipedia dumps.

For evaluation, we adopt two metrics, Precision@N and Mean Average Precision (MAP) to quantify the performance of different methods. In this experiment, we adopt six methods as baselines. For detailed information about the baseline methods, please refer to (Tang et al., 2012).

5.2 Experimental Results

Table 1 shows the experiment results of different methods. From Table 1, we can see that the proposed approach outperforms all the baselines. Through analyzing the experiments, we find anchor probability is very efficient in key term selection, since it could filter out most unimportant key term candidates. Topical relevance and key term normalization could also improve the performance. Knowledge graph based method translation could get high precision results, and machine translation system could provide complementary information for knowledge graph based translation.

Method	MAP	P@5	P@10	P@20
CELD	**0.217**	0.767	**0.733**	**0.653**
LinkProb	0.168	**0.800**	0.694	0.546
PNM	0.123	0.667	0.567	0.499
LinkProbEn2	0.095	0.456	0.428	0.338
LinkProbEn	0.085	0.489	0.394	0.315
LinkProb_S	0.059	0.411	0.322	0.268
LinkProbEn_S	0.033	0.233	0.186	0.144

Table 1. Experiment results

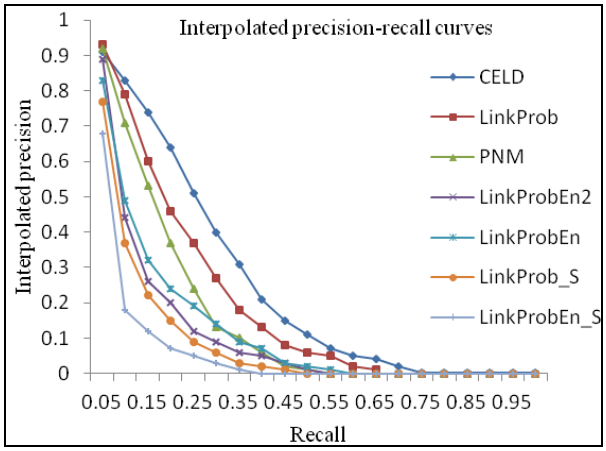

Figure 5. The precision/recall curves of CELD system

Figure 5 shows the interpolated precision-recall curves of CELD and other baseline methods. From Figure 5, we can see the proposed system outperforms all the baseline methods.

6 Conclusion

In this paper we present a hybrid approach for Chinese to English link discovery. This approach can automatically identify anchors in Chinese document and link to target concepts in English Wikipedia. To solve the Chinese character variant issues, we develop a normalization lexicon. We also build a knowledge graph for key term translation. Experimental results on real world datasets show promising results and demonstrate the proposed approach is efficient. As a future research, we plan to use more sophisticated nature language processing techniques to key term extraction and translation. We also plan to integrating linking and contextual information for concept resolution.

[8] http://crosslink.googlecode.com/files/zh-topics-36.zip

References

Bontcheva, K., and Rout, D. 2012. Making Sense of Social Media Streams through Semantics: a Survey. Semantic Web journal.

Cassidy, T., Ji, H., Deng, H. B., Zheng, J., and Han, J. W. 2012. Analysis and Refinement of Cross-Lingual Entity Linking. In Proceedings of the third International Conference on Information Access Evaluation: Multilinguality, Multimodality, and Visual Analytics,2012. CLEF'12. Springer-Verlag Berlin, Heidelberg, 1-12.

Fahrni, A., Nastase, V., and Strube, M., 2011. HITS' Graph-based System at the NTCIR-9 Cross-lingual Link Discovery Task. In Proceedings of ntcir-9 workshop meeting, 2011. NTCIR'9.

Han, X. P., and Sun, L. 2012. An Entity-Topic Model for Entity Linking. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 105-115.

Kim, J., and Gurevych, I. 2011. UKP at CrossLink: Anchor Text Translation for Cross-lingual Link Discovery. In Proceedings of ntcir-9 workshop meeting, 2011. NTCIR'9.

Kiryakov, A., Popov, B., Ognyanoff, D., Manov, D., Kirilov, A., and Goranov, M. 2004. Semantic Annotation, Indexing and Retrieval. Journal of Web Semantics, ISWC 2003 Special Issue, 1(2): 49-79, 2004.

Mihalcea, R., and Csomai, A. 2007. Wikify! Linking Documents to Encyclopedic Knowledge. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, 2007, CIKM'07. ACM New York, NY, 233-242.

Milne, D., and Witten, I. H. 2008. Learning to link with Wikipedia. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, 2008. CIKM'08. ACM New York, NY, 509-518.

Tang, L. X., Geva, S., Trotman, A., Xu, Y., and Itakura, K. Y. 2011. Overview of the NTCIR-9 Cross-link Task: Cross-lingual Link Discovery. In Proceedings of ntcir-9 workshop meeting, 2011. NTCIR'9.

Tang, L. X., Trotman, A., Geva, S., and Xu, Y. 2012. Cross-Lingual Knowledge Discovery: Chinese-to-English Article Linking in Wikipedia. Lecture Notes in Computer Science Volume 7675, 2012, 286-295.

Wang, Z. C., Li, J. Z., Wang, Z. G., and Tang, J.2012. Cross-lingual Knowledge Linking across Wiki Knowledge Bases. In Proceedings of the 21st International Conference on World Wide Web, 2012. WWW'12. ACM New York, NY, 459-468.