Knowledge-Based Systems 111 (2016) 228–236

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Cross-language article linking with different knowledge bases using bilingual topic model and translation features

Yu-Chun Wang[a], Chun-Kai Wu[b], Richard Tzong-Han Tsai[c,d,*]

[a] Telecommunication Laboratories, Chunghwa Telecom, Taiwan
[b] Department of Computer Science, National Tsinghua University, Taiwan
[c] Department of Computer Science and Information Engineering, National Central University, Taiwan
[d] NTU IoX Center

ARTICLE INFO

Article history:
Received 8 October 2014
Revised 8 August 2016
Accepted 11 August 2016
Available online 17 August 2016

Keywords:
Cross-language article linking
Link discovery
Bilingual topic model

ABSTRACT

Creating links among online encyclopedia articles in different languages is crucial in the construction and integration of large multilingual knowledge bases. Most research to date has focused on linking among different language versions of Wikipedia, yet other large online encyclopedias in a variety of languages exist. In this work, we present a cross-language article-linking method using a bilingual topic model and translation features based on an SVM model to link articles in English Wikipedia and Chinese Baidu Baike, the most widely used Wiki-like encyclopedia in China. To evaluate our approach, we compile data sets from Baidu Baike articles and their corresponding English Wikipedia articles. The evaluation results show that our approach achieves at most 0.8158 in MRR, outperforming the baseline system by 0.1328 (+19.44%) in MRR. Our method does not heavily depend on linguistic characteristics, and it can be easily extended to generate cross-language article links among different online encyclopedias in other languages.

1. Introduction

Online encyclopedias are a rich source of human knowledge, and linking related articles among online encyclopedias in different languages is crucial for global knowledge sharing. Cross-language article linking (CLAL) is the task of creating links between encyclopedia articles in different languages that describe the same content. CLAL has applications in many research fields such as named entity translation [1], cross-language information retrieval [2], and multilingual knowledge base creation [3].

Much CLAL work has been done on Wikipedia. With more than 34 million articles in 287 languages, Wikipedia is the largest multilingual online encyclopedia. It is a collaborative encyclopedia with more than 35 million volunteers from all around the world. Articles in Wikipedia are partly structured, usually containing a title, table of contents, main context, related images, audio or video files, categories, and infoboxes (structured table of metadata). Wikipedia articles may also have inter-language links to corresponding articles in other language versions. However, these links must first be made by human users, so some articles lack inter-language links. Several approaches have been proposed to automatically generate inter-language links between different language versions of Wikipedia [4,5].

Although Wikipedia is one of the largest online encyclopedias and a useful multilingual knowledge base, the coverage of different language editions is inconsistent. For instance, there are 4.7 million English articles, but only 800,000 articles in Chinese and 300,000 in Korean. Of course, inter-language links cannot be created to non-existent pages.

On the other hand, in some languages, such as Chinese and Korean, alternate large monolingual knowledge bases exist. For example, Baidu Baike, it is an online collaborative Wiki-like encyclopedia provided by the Chinese search engine company, Baidu. Started in 2006, it now contains more than 10 million Chinese articles and has about 5 million registered users. Like Wikipedia, Baidu Baike is an open encyclopedia with topics covering most branches of knowledge. Baidu Baike also contains several article types that are not permitted on Wikipedia, such as word definitions, recipes, and commercial product pages. The articles in Baidu Baike are all in simplified Chinese, the most commonly used text script in mainland China. Articles structure in Baidu Baike is similar to that in Wikipedia. The articles usually contain a title, table of contents, main context, images, and hyperlinks to other articles. Baidu Baike also has several pre-defined fix-format tables,

* Corresponding author.
E-mail addresses: ycwang@cht.com.tw (Y.-C. Wang), s102065512@m102.nthu.edu.tw (C.-K. Wu), thtsai@csie.ncu.edu.tw (R.T.-H. Tsai).

http://dx.doi.org/10.1016/j.knosys.2016.08.015
0950-7051/© 2016 Elsevier B.V. All rights reserved.

similar to Wikipedia's infoboxes. However, unlike infoboxes, these fix-format tables cannot be freely added or modified by ordinary users.

Besides Baidu Baike, there are several other large online encyclopedias in China with similar numbers of articles and users. For example, Hudong Baike (Alexa daily pageviews: 8,736,467) has 8 million articles and 5 million registered users. Sogou Baike is another newer competitor. We chose Baidu Baike as our dataset because it remains the most widely used (Alexa daily pageviews: 65,080,009), partly because Baidu is the number 1 search engine in China, and partly because Baidu tends to link to its own encyclopedia rather than Hudong Baike [6,7].

When alternatives to Wikipedia exist in multiple languages, researchers can enrich or broaden their data resources by creating cross-language links among different knowledge bases, building meta-encyclopedias that include many online encyclopedias in different languages. The main challenge in this task is that there is no common linking information across knowledge bases and article format and structure differ. This means that much of the previous work done on Wikipedia (see Section 2) cannot be applied to CLAL among different knowledge bases. Therefore, a new approach is required. In this paper, we propose a CLAL method to select and link related articles between the English Wikipedia and Baidu Baike online encyclopedias. In the field of CLAL, our approach is the first bilingual topic modeling method that can discover hidden topics in bilingual encyclopedia articles. In addition, we propose several novel translation features, including hypernym extraction, infobox similarity, and English title occurrence, to improve performance. Our other major contribution to CLAL research is the compilation of three benchmark datasets from the two encyclopedias above to evaluate the linking accuracy of CLAL.

Once we have developed a CLAL system our next goal is the construction of a meta-encyclopedia.[1] There are several applications and benefits of CLAL. For example, after linking articles among different knowledge bases, we can regard corresponding linked pairs as translations of each other (at least in part). This can be very useful in building named entity dictionaries, which are essential in the difficult task of translating named entities. Another application of CLAL is cross-language/cultural analysis of concepts and entities. The articles in online encyclopedias are collaboratively created by a wide user base and thus reflect common knowledge in a language/culture. For example, Chinese encyclopedias usually contain very detailed articles about famous Chinese entertainers, while English encyclopedias may have only cursory information or lack of such articles entirely. After linking corresponding articles, we can statistically analyze cross-language/cultural similarity of certain concepts. In addition, CLAL also provides the opportunity to observe the borrowing of new concepts from one language/culture into another. We can perform CLAL at different time periods and examine how cross-language links develop over time. Thus CLAL may also find applications in social science research fields.

2. Related work

Cross-language article linking between different encyclopedias is a new research target, and there is little previous research in this field. However, several tasks in natural language processing are closely related to cross-language article linking, including entity linking, cross-language entity linking, and missing link discovery in Wikipedia. In the following section, we summarize the research into these related tasks.

Entity linking (EL) is the task of determining the identities of entities mentioned in text and linking the entities to entries in a knowledge base (KB). Wikipedia is the most widely used knowledge base in entity linking research. One of the most important organized shared tasks of EL is the Knowledge Base Population (KBP) track at the NIST Text Analysis Conference (TAC) [8,9]. In the first TAC-KBP, KBP 2009, the EL task was mono-lingual English EL with Wikipedia as the reference KB. The input query was an English entity (person name, location, or organization) and a news document containing this entity. Participant systems had to determine the correct corresponding KB entry for the given entity.

EL approaches can be divided into three types: rule-based, supervised-learning-based, and graph-model-based. Guo et al. [10] proposed a rule-based method to score the KB candidates based on textual similarity. They used the Dice coefficient, string edit distance and the Jaccard similarity coefficient to measure the similarity of the given entity (within the context in which it appears) and any KB entry, and they then combined these three similarity scores linearly into a final score. Zhang et al. [11] used a supervised ranking SVM model to rank the KB entries. They designed several features to measure similarity between the given entity and a KB entry candidate, including name string similarity, TF-IDF document similarity, and topic similarity. Fahrni et al. [12] proposed an EL method based on the Markov logic network model. They defined several first-order logic formulae to describe the relatedness and similarity between entities and KB entries. Hachey et al. [13] proposed a graph-based EL method. They first constructed a graph composed of all Wikipedia KB entry candidates. Each vertex in the graph corresponds to a Wikipedia article, and each directed edge represents a hyperlink from one article to another article. They then used graph measures such as degree centrality and PageRank to rank KB entry candidates.

The EL task was then extended to the cross-language entity linking (CLEL) task, in which the language of the given entity is different from that of the reference KB. McNamee et al. [14] compiled a multilingual test set for CLEL from the TAC-KBP 2009 dataset by translating the dataset queries into 21 languages and collecting new documents in those languages. The KB they used was English Wikipedia. They proposed a cross-language entity-linking approach with two steps: candidate selection and candidate ranking. In candidate selection, they transliterated the queries into English and then selected possible candidates based on English heuristics such as exact match, alias or nickname, and common 4-gram characters. In the second step, they ranked each candidate using an SVM model (SVM^{rank}) with many NLP textual features such as contextual similarity, relation features, and entity type features.

CLEL systems usually employ translation or transliteration methods to process multiple languages. Clarke et al. [15] used LDC's Chinese-English Named Entity Lists which contain over 500,000 name pairs, and the Rosette Name Translator. Miao et al. [16] adopted a statistical hierarchical phrase-based machine translation method in their Chinese-English EL system.

The main difference between CLAL and (CL)EL is that in CLAL the article in which the given entity appears describes that entity instead of merely mentioning it. In other words, the query document in EL or CLEL is just a text string containing the entity, and the surrounding text in the document may not directly describe the entity mention; while in CLAL, the original article and the linked article should both contain detailed information about the given entity as well as related structural information.

The last related task we will discuss is cross-language link discovery (CLLD), which aims to create missing cross-language links among articles in different Wikipedia language versions. Sorg and Cimiano [4] constructed an SVM-based classifier with graph-based and text-based features to predict whether pairs of English and German Wikipedia articles should be linked or not. The graph features are the counts of chain links, the similarity of chain links, and common categories. The text features are the string edit distance

[1] This research is being carried out at National Central University, Taiwan as part a three-year project funded by the Taiwanese Ministry of Science and Technology.