Expanding English and Chinese Dictionaries by Wikipedia Titles

Wei-Ting Chen, Yu-Te Wang and Chuan-Jie Lin
Department of Computer Science and Engineering
National Taiwan Ocean University
{10757025, 00557025, cjlin}@ntou.edu.tw

Abstract

This paper introduces our preliminary work in dictionary expansion by adding English and Chinese Wikipedia titles along with their linguistic features. Parts-of-speech of Chinese titles are determined by the majority of heads of their Wikipedia categories. Proper noun detection in English Wikipedia is done by checking the capitalization of the titles in the content of the articles. Title alternatives will be detected beforehand. Chinese proper noun detection is done via interlanguage links and POS. The estimated accuracy of POS determination is 71.67% and the accuracy of proper noun detection is about 83.32%.

Keywords—dictionary expansion, proper nouns, parts-of-speech, Wikipedia

1 Introduction

Dictionaries play an important role in many NLP researches. A dictionary contains a list of words. It can be used to provide candidates in Chinese word segmentation. If a dictionary also collect phrases, it can help to detect syntactic units when doing syntax parsing. Some dictionaries provide information about parts-of-speech or semantics, which is important for POS tagging and many other NLP applications (Harabagiu and Hickl, 2006; Allam and Haggag, 2012; Liu *et al.*, 2016).

A major issue of using dictionaries is the expansion of unknown words. This issue is especially important in Chinese word segmentation. If the dictionary does not recognize many new words, it is impossible to segment an input sentence correctly. It might be easy to collect unknown words from the Internet such as Wikipedia, but their parts-of-speech or other important linguistic features are not easy to be determined, because their sources are not designed for NLP purposes.

In NLP domain, there have been many researches about extracting information from Wikipedia in different aspects and methods. Popular researches include knowledge base expansion (Ji and Grishman, 2011), Wikipedia article similarity measurement by the hierarchy of categories (Ponzetto and Strube, 2007; Witten and Milne, 2008), infobox completion (Wu and Weld, 2008), and so on. Many NLP applications used Wikipedia as a resource, such as improving machine translation by Wikipedia interlanguage links (Jones *et al.*, 2008; Nguyen *et al.*, 2009), measuring document similarity (Nakamura *et al.*, 2014), word sense disambiguation (Hoffart *et al.*, 2011), annotating Wikipedia entries in documents (Kulkarni *et al.*, 2009), and question answering (Buscaldi and Rosso, 2006; Waltinger *et al.*, 2011), including answer-type decision by Wikipedia (Huang *et al.*, 2008; Chen *et al.*, 2016).

This paper proposes methods to expand English and Chinese dictionaries by adding titles of Wikipedia articles, for they are new and constantly maintained. Note that many of them are indeed multi-word phrases. Methods to add linguistic features to these new words, such as parts-of-speech, and proper nouns, are also discussed.

This paper is organized as follows. Section 2 describes how we preprocess data in Wikipedia. Section 3 introduces our approach to determine parts-of-speech. Section 4 proposes methods to decide whether a title is a proper noun or not. Section 5 shows the experimental results and Section 6 concludes the paper.

2 Preprocessing Wikipedia Data

Wikipedia is a collaborative encyclopedia contributed by real users around the world. It contains millions of Chinese pages and tens of millions of English pages. However, not all pages are main encyclopedia articles, such as the pages listed as follows. They should be discarded beforehand.

(1) **Removing administration pages:** An administrative page usually has a title containing a semicolon but not followed by a whitespace, such as “Help:Category” in English Wikipedia or “使用說明:分類” in Chinese Wikipedia. These are not main articles hence should be removed.

(2) **Removing disambiguation pages:** A disambiguation page defines an ambiguous term by providing a list of Wikipedia articles as reference. For example, the term “blue” may refer to a color “Blue”, a movie “Three Colors: Blue”, or many other meanings provided in the page “Blue (disambiguation)”. These are not main articles hence should be removed.

(3) **Removing list pages:** A list page provides a list of Wikipedia articles concerning a specific topic, such as “List of Game of Thrones characters”. We think the titles of these pages are not suitable for dictionaries and discard them. The title of an English list page usually starts with “List of” or “Lists of”. A Chinese list page is often categorized under “xxx 索引” (*index*) or “xxx 列表” (*list*).

(4) **Pending redirect pages:** Redirect pages often link titles to an authoritative and highly correlated page, sometimes these titles are paraphrases. For example, both “US” and “USA” redirect to the page “United States”. These pages are not included in our work described in this paper. In the future, if we can determine whether the title of a redirect page is a paraphrase to a main page, we can add this title into our dictionary and share the same linguistic features as its authoritative title.

After removing these pages, all other titles will be added into our dictionaries. The following sections propose methods to decide their linguistic features.

3 Parts-of-Speech Determination

Our first challenge is to determine the part-of-speech of each Wikipedia title. POS information is not explicitly given in Wikipedia but is essential for many NLP techniques.

The solution might be easier in English, because it is a convention to write the title of an encyclopedia article in its nominal form. For example, the editors tend to create an article named “humiliation” rather than “to humiliate”. However, its corresponding Chinese title “羞辱” is a verb.

Since most of the English titles are nouns, we can determine the part-of-speech of a title by deciding whether this title is in singular or plural form (NN vs. NNS) or a common noun or proper noun (NN vs. NNP). This process is not the main focus of this paper thus will not be discussed here.

Chinese Wikipedia titles may have parts-of-speech other than nouns. In Chinese, many verbs and adjectives can be nominalized without inflection, such as the previous example “羞辱”. These words only have parts-of-speech of verbs in Academia Sinica Lexicon[1], a standard Chinese lexicon, and will be not tagged as Nv (nominalized verb) until being nominalized in sentences. Therefore, there are many possible POS candidates for Chinese titles.

Besides, as defined in the Academia Sinica Lexicon, there are 44 different parts-of-speech in Chinese. We only consider open classes, i.e. nouns, verbs, and adjectives, to be POS candidates of new terms. A complete list is shown in Table 1.

POS	Meaning
A	Non-predicative adjective
Na	Common Noun
Nb	Proper Noun
Nc	Place Noun
Nd	Time Noun
VA	Active Intransitive Verb
VAC	Active Causative Verb
VB	Active Pseudo-transitive Verb
VC	Active Transitive Verb
VCL	Active Verb with a Locative Object
VD	Ditransitive Verb
VE	Active Verb with a Sentential Object
VF	Active Verb with a Verbal Object
VG	Classificatory Verb
VH	Stative Intransitive Verb
VHC	Stative Causative Verb
VI	Stative Pseudo-transitive Verb
VJ	Stative Transitive Verb
VK	Stative Verb with a Sentential Object
VL	Stative Verb with a Verbal Object

Table 1: Open Classes of Chinese Parts-of-Speech.

[1] http://ckip.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm

Our approach to decide a Chinese title's POS is counting the parts-of-speech of the heads of the title's categories. As another convention, Wikipedia categories are often written in noun phrases, too. The heads of these phrases are often (but not always) hypernyms of the Wikipedia title. For example, the term "自燃" (*spontaneous combustion*) belongs to the category "燃燒" (*burning*) and hence is a verb in Chinese. Similarly, the term "自由滑" (*free skating*) is a verb because it belongs to the category "花式滑冰" (*figure skating*) whose head "滑冰" (*skating*) is a verb in Chinese.

Since Chinese is a head-final language, i.e. the head of a phrase appears at the end of that phrase, we segmented the Wikipedia category labels by the Academia Sinica Lexicon and treated the last words as the heads.

The same approach might be able to predict a title's semantic class as well, because the most frequent head is highly related to its semantic class. Take the term "飆風特攻" (Point Break) as an example. Heads of 16 of its Chinese categories are "電影" (*movie*). 6 of them are hyponyms of "movie", including "驚悚片" (*thriller film*) and "動作片" (*action movie*). Only 2 heads are not related to "movie". We can say that the semantic class of "飆風特攻" is "movie". As we have also observed some counter examples, this idea will be explored more fully in the future.

4 Proper Noun Identification

Our second challenge is to identify proper nouns from these Wikipedia titles. Some Wikipedia articles are about proper nouns (such as "United Nations") and some are not (such as "Rainbow"). As we know that proper nouns carry more information than common nouns, it will be great if the information of proper nouns can be added into a dictionary.

Proper nouns can be identified based on two features. One feature is infobox template or category type. The other is the distribution of capitalization. We will discuss these two features in the following sections.

4.1 Finding Proper Nouns by Special Categories

Some classes of Wikipedia articles can be easily decided by their categories and infoboxes. We use the following rules to collect articles in special categories and decide their properness.

(1) **Person names:** If an article uses an infobox template which belongs to the category of "people and person infobox templates" (in English Wikipedia) or "人物信息框模板" (in Chinese Wikipedia), its title is considered as a person name.

(2) **Location names:** If an article uses an infobox template which belongs to the category of "geography and place infobox templates" (in English Wikipedia) or "地理和場所信息框模板" (in Chinese Wikipedia), its title is considered as a location name.

(3) **Letters and numbers:** Pages about alphabets of a language, numbers, dates, and domain names are NOT proper nouns. They belong to some specific categories and can be easily identified.

When we say that an article belongs to Category B, we mean that this article belongs to a category that is Category B itself or B's descendant in the Wikipedia category hierarchy.

4.2 Finding Proper Nouns by Capitalization

Proper nouns in English are usually capitalized. However, it is not that straightforward because a) titles of Wikipedia articles do not always appear in the content, and b) words at the beginning of sentences are also capitalized.

To deal with the first issue, we find that the absent titles may be mentioned by using different words or phrases. These alternatives can be found by several ways as discussed later.

To identify proper-noun titles, we define a set of rules to classify them into several groups. The main feature in the grouping rules is the frequency of a title being capitalized in its article. The majority in a group defines its likelihood of having proper nouns as members.

Title Alternatives

In the case that a title does not appear in the article by itself, we need to find its alternatives by the following methods before counting the frequency of capitalized cases.

(1) Removing phrases after punctuations
Sometimes an additional phrase is attached to a title with a punctuation for the purpose of

disambiguation. Such a phrase is not part of a name and should be removed.

> **Arena, North Dakota** (*alternative: "Arena"*)
> Arena is an extinct town in Burleigh County, North Dakota. The GNIS classifies it as…

(2) Removing parenthesized phrases
Sometimes a parenthesized string is attached to a title for the purpose of disambiguation. Such a phrase is not part of a name and should be removed.

> **Android (robot)** (*alternative: "Android"*)
> An android is a humanoid robot or synthetic organism designed to look and act like…

(3) Detecting boldfaced phrases in the lead sections
A lead section[2] of a Wikipedia article "*serves as an introduction to the article and a summary of its most important contents*." If the author wants to use a different string to substitute the title, he or she will introduce this phrase in the lead section in boldface.

> **Namhansan** (*alternative: "Namhan Mountain"*)
> **Namhan Mountain** is a 460 m peak in Gyeonggi-do province, South Korea…

If there are two or more boldfaced phrases in the lead section, we choose the one which appears the most times in the article, or the most similar one (by edit distance) if tied. In the following example, the title "Halahala" does not appear in the article at all, but "Halāhala" appears many times, hence being chosen as the title's alternative.

> **Halahala** (*alternative: "Halāhala"*)
> **Halāhala** (Sanskrit हलाहल) or **kālakūṭa** (Sanskrit कालकूटं, literally: 'black mass' or 'time puzzle') is the name of a poison…

Grouping Rules

For each title (or its alternative if the original title does not appear in the article), we first count the times it is being capitalized or lowercased, at the beginning or in the middle of a sentence. But it does not count if the target title appears in a longer Wikipedia title. For example, the article of the term "football" contains the sentence "*In 1888, the Football League was founded in England, becoming the first of many professional football competitions.*" The capitalized phrase "*the Football League*" is another Wikipedia title. So the term "football" only counts once, as "being lowercased in the middle of a sentence".

According to the times of a title being capitalized in its article, we define 10 groups as follows. The grouping rules are applied in order. One title can only belong to one group. A title which does not match any rule will remain unclassified.

G00: If a title has one capital letter which is not at its beginning, such as the capital 'M' in the title "HiM", it belongs to this group and is considered as a proper noun. Note that a title only being uppercased in its article also belongs here. An example is the title "Lindauer Dornier" in the sentence "*Lindauer DORNIER GmbH is a family-owned business…*"

G01: If a single-word title has ever been capitalized inside (not at the beginning of) a sentence, such as "Animalia" in the sentence "*Julia MacRae Books published an Animalia colouring book in 2008*", it belongs to this group and is considered as a proper noun.

G02: If a single-word title is more often lowercased than capitalized inside its article, it belongs to this group and is considered as a common noun. An example is the title "Chuckwagon" which is found 4 times capitalized but 20 times lowercased inside its article. It is a type of wagon so it is just a common noun.

G03: If a single-word title is only lowercased in its article, it belongs to this group and is considered as a common noun.

G04: If a single-word title is only capitalized in its article but all at the beginning of sentences, it belongs to this group. Following the majority rule, i.e. most of the observed titles in this group are proper nouns, they are considered as proper nouns.

G05: If a single-word title does not appear in its article at all, it belongs to this group.

[2] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section

Following the majority rule, they are considered as common nouns.

G06: If a multi-word title contains a lowercased content word, it belongs to this group and is considered as a common noun. An example is the title “Ames test”, where “test” is a content word but lowercased.

G07: If a multi-word title is capitalized at least once in its article, such as the title “First Women's Bank of California” in the sentence “*The First Women's Bank of California was a local Los Angeles bank dedicated to…*”, it belongs to this group and is considered as a proper noun.

G08: If a multi-word title appears in its article but written in lowercase, such as the title “Push Stick” in the sentence “*The purpose of a push stick is to help the user…*”, it belongs to this group and is considered as a common noun.

G09: A multi-word title does not appear in its article, but its head is often capitalized inside the article, it is considered as a proper noun.

4.3 Finding Proper Nouns by Interlanguage Links and POS

Because Chinese does not have capitalization, our first approach to identify Chinese proper nouns is via interlanguage links which connect Wikipedia articles discussing the same topic but in different languages. If a Chinese article links to an English article in the groups G00 ~ G09, the Chinese title joins that group and can be classified as its majority.

However, half of the Chinese articles do not link to any English articles. We need to look for other solutions in the future.

5 Experiments

The experimental data are dumped files of English Wikipedia and Chinese Wikipedia on 2019/2/1. After removing administrative, disambiguation, and list pages, there are totally 5,679,503 English Wikipedia main articles and 995,294 Chinese Wikipedia main articles. After further identifying person names, location names, letters and numbers, 3,905,050 English and 839,174 Chinese articles are waited to be processed as shown in Table 2.

5.1 Parts-of-Speech Determination for Chinese Titles

To observe the correctness of POS determination, we chose 21,211 Chinese titles which have appeared and been manually POS-tagged in ASBC (Academia Sinica Balanced Corpus) (Chen *et al.*, 1996) and have category labels in Wikipedia. Note that they may have more than one POS in ASBC. Methods introduced in Section 3 were used to predict POS sets for these Chinese titles.

We made three different observations. If only top 1 choice counted, there were 11,459 correct guessing with an accuracy of 54%. If any match in the POS sets counted, 15,201 correct guessing made an accuracy of 71.67%. It means that the guessed POS sets were efficient candidates for further verification in the future.

One major type of the errors is that categories tend to be nouns. For example, the term “羞辱” (humiliation) is a verb in Chinese, but its categories are “道德” (morality), “情緒” (emotions), and “性行為” (sexual acts), which are all nouns in Chinese.

5.2 Proper Noun Detection in English

3,905,050 English main articles were grouped by the rules introduced in Section 4. There were still 125,924 articles left as unclassified.

To evaluate the accuracy of proper noun detection, we randomly selected a small subset in

Type	English	Chinese
Main pages	5,679,503	995,294
Person names	1,505,094	49,592
Location names	263,805	100,586
Letters & numbers	5,554	5,942
To be processed	3,905,050	839,174

Table 2: Pre-Determined Wikipedia Pages.

Group	Articles	Proper	Subset	Acc.
G00	330,569	Y	101	84.16%
G01	450,022	Y	107	80.37%
G02	22,947	N	103	100.0%
G03	24,490	N	103	96.12%
G04	150,090	Y	110	91.82%
G05	46,764	N	103	56.31%
G06	957,198	N	162	82.72%
G07	1,581,486	Y	222	83.33%
G08	470	N	107	83.18%
G09	238,037	Y	108	84.62%
noG	125,924	N	108	83.33%

Table 3: English Proper Noun Detection Accuracy.

each group and assessed by human. The performance is shown in Table 3. The estimated overall accuracy is about 83.32%.

One type of the errors is that a common-noun title appears inside a proper name but the proper name is not a Wikipedia title, such as “chuckwagon” in the name “the American Chuckwagon Association”. If the title does not appear many times, it will be misclassified as a proper noun.

Another type of the errors is the inconsistent capitalization of the writers. For example, the term “East Bradford Township” is the name of a township in US, but the authors of this Wikipedia article also use “East Bradford township” to refer to this area, and the term is misclassified as a common noun.

5.3 Proper Noun Detection in Chinese

Only 554,892 of 839,174 Chinese articles can be mapped to English articles via interlanguage links. 85,302 of them map to person names, 42,727 to location names, and 4,323 to letters or numbers. 410,258 are grouped into G00 ~ G09 and 292,268 of them are classified as proper nouns.

6 Conclusion

This paper depicts our preliminary work of dictionary expansion by adding Wikipedia titles. The parts-of-speech of Chinese titles are decided by the voting of heads of their categories. English proper nouns are identified by cases of capitalization, while Chinese ones are identified via interlanguage links. These methods achieved an accuracy of 71.67% in POS determination and 83.32% in English proper noun detection.

More modern methods will be experimented in the future. Sentences containing Wikipedia titles will be POS-tagged to decide the titles' POS. Their semantic classes and properness can be determined by deep learning.

Acknowledgments

This research was funded by the Taiwan Ministry of Science and Technology (grant: MOST 106-2221-E-019-072.)

References

Ali Mohamed Nabil Allam and Mohamed Hassan Haggag. 2012. The question answering systems: a survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3):211-221.

David Buscaldi and Paolo Rosso. 2006. Mining knowledge from Wikipedia for the question answering task. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 727-730.

Keh-Jiann Chen, Chu-Ren Huang, Li-ping Chang, and Hui-Li Hsu. 1996. Sinica Corpus: design methodology for balanced corpora. In *Proceeding of the 11th Pacific Asia Conference on Language, Information and Computation*, pages 167-176.

Po-Chun Chen, Meng-Jie Zhuang, and Chuan-Jie Lin. 2016. Using Wikipedia and semantic resources to find answer types and appropriate answer candidate sets in question answering. In *Proceedings of the Open Knowledge Base and Question Answering (OKBQA) Workshop, the 26th International Conference on Computational Linguistics (COLING 2016)*, pages 1-10.

Sanda Harabagiu and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 905-912.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP ‘11)*, pages 782-792.

Zhiheng Huang, Marcus Thint, and Zengchang Qin. 2008. Question classification using head words and their hypernyms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP ‘08)*, pages 927-936.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT ‘11)*, pages 1148-1158.

Gareth J.F. Jones, Fabio Fantino, Eamonn Newman, and Ying Zhang. 2008. Domain-specific query translation for multilingual information access using machine translation augmented with dictionaries mined from Wikipedia. In *Proceedings of CLIA 2008 - 2nd International Workshop on Cross Lingual Information Access, the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pages 34-41.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of Wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining (KDD '09), pages 457-466.

Chi-Ting Liu, Shao-Heng Chen, and Chuan-Jie Lin. 2016. Learning textual entailment classification from a Chinese RITE dataset specialized for linguistic phenomena. In *Proceedings of IEEE 17th International Conference on Information Reuse and Integration (IRI), IEEE International Workshop on Empirical Methods for Recognizing Inference in Text (EMRITE)*, pages 506-512.

Tatsuya Nakamura, Masumi Shirakawa, Takahiro Hara, and Shojiro Nishio. 2014. Semantic similarity measurements for multi-lingual short texts using Wikipedia. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pages 22-29.

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. *Evaluating Systems for Multilingual and Multimodal Information Access*, 5706:58-65.

Simone Paolo Ponzetto and Michael Strube. 2007. Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research (JAIR)*, 30:181-212.

Ulli Waltinger, Alexa Breuing, and Ipke Wachsmuth. 2011. Interfacing virtual agents with collaborative knowledge: open domain question answering using Wikipedia-based topic models. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-11)*, pages 1896-1902.

Ian Witten and David Milne. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pages 25-30.

Fei Wu and Daniel S. Weld. 2008. Automatically refining the Wikipedia infobox ontology. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*, pages 635-644.