Towards Cross-lingual Patent Wikification

Takashi Tsunakawa tuna@inf.shizuoka.ac.jp
Hiroyuki Kaji kaji@inf.shizuoka.ac.jp
College of Informatics, Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka 432-8011, Japan

Abstract

This paper demonstrates the effectiveness of cross-lingual patent wikification, which links technical terms in a patent application document to their corresponding Wikipedia articles in different languages. The number of links increases definitely because different language versions of Wikipedia cover different sets of technical terms. We present an experiment of Japanese-to-English cross-lingual anchor text extraction using a dedicated technical term extraction system and a patent parallel corpus. Cross-lingual anchor text extraction retrieves about 10% more technical terms linked to Wikipedia articles than monolingual extraction. We also show that restricting anchor texts to technical terms in a specified Wikipedia category has effect of reducing the number of destination article candidates.

1. Introduction

Wikification refers to a task of linking phrases in a text to their corresponding Wikipedia articles. It greatly enhances the understandability of the text; namely, readers of a wikified text can easily figure out the meaning of an unfamiliar phrase by clicking it. Applying wikification to patent application documents is helpful for understanding technical terms in them. With the rapidly increasing quantity and improved quality of Wikipedia articles in technical domains, the effectiveness of patent wikification will be enhanced.

A promising extension of wikification is cross-lingual wikification, which links phrases in a text to articles in languages other than that of the text. Wikipedia has more than 35 million articles in more than 290 languages. Since each language version is being edited independently, the quantity and quality of articles are very diverse among languages. While English Wikipedia has nearly five million articles, most other language versions have less than one million articles as of 2015. This indicates that a huge number of entities in the world are explained only in English. Thus, we can enrich wikification results by considering links from other languages to English. Consider a Japanese patent application document that includes the technical term "アーク抑制 (*aku yokusei*)," which is not explained by any Japanese Wikipedia article but by the English Wikipedia article "Arc suppression." Cross-lingual wikification links the technical term "アーク抑制 (*aku yokusei*)" to the English Wikipedia article "Arc suppression." Such cross-lingual links would be very useful for revealing important entities that could not be found by monolingual wikification.

In this paper, we demonstrate the potential effectiveness of cross-lingual patent wikification. The main target is patent application documents in languages other than English (such as Chinese and Japanese), which have been increasing by leaps and bounds. We showed in an experiment that cross-lingual wikification to Japanese patent application documents could significantly increase the number of links by adding English Wikipedia as the linking destination.

2. Related work

Wikification consists of two steps: anchor text extraction and disambiguation to Wikipedia articles. We describe studies on the former step that is the main concern of this paper. Anchor texts are phrases that should be linked to Wikipedia articles. One of the most important types of information for anchor text extraction is keyphraseness, namely link probabilities that phrases are used as anchor texts for linking to Wikipedia articles (Mihalcea and Csomai, 2007). Another is relatedness with co-occurring phrases (Milne and Witten, 2008) because a text tends to be linked to articles related to it. Keyword extraction techniques are also applicable to anchor text extraction, because keywords in a text should be anchor texts (Mihalcea and Csomai, 2007). There have been many studies for keyword extraction using syntactic and statistical information (Jacquemin and Bourigault, 2003).

With ever more attention being focused on wikification, cross-lingual wikification has been recognized as a new challenging task. Cross-lingual wikification consists of anchor text extraction, translation, and disambiguation. Translation quality severely affects the wikification results with this approach. To our knowledge, translation of technical terms for cross-lingual wikification has not been studied. In the previous studies, a large part of anchor texts are named entities, translation of which requires special techniques such as transliteration, name translation mining from comparable corpora, and information extraction-based techniques (McNamee et al., 2011; Cassidy et al., 2012; Miao et al., 2013).

3. Monolingual vs. cross-lingual anchor text extraction

In this section, we give a detailed description of our approach to cross-lingual patent wikification. Since our present goal is to estimate increases in the number of links by introducing cross-lingual wikification, we focus mainly on the anchor text extraction problem.

What kinds of phrases should be extracted as anchor texts depends on the domain of the text and the application of wikification results. Most anchor texts to be extracted in patent wikification are technical terms, while original wikification (Mihalcea and Csomai, 2007) also extracts named entities. Named entities, such as personal names and place names, seldom occur in patent application documents. In most cases, these are of no interest to readers for understanding what is invented in the patent. We therefore extract technical terms from patent application documents as anchor texts. In order to discriminate technical terms from other kinds of anchor texts, we employ a technical term extraction system.

Our anchor text extraction system built for an experiment consists of three parts: technical term extraction and monolingual/cross-lingual anchor text extraction (Fig. 1).

1) Technical term extraction

We first extract technical terms from patent application documents as anchor candidates by a technical term extraction system. To extract terms from Japanese patent application documents, we employ *termex*,[1] the automatic domain terminology extraction system developed at Nakagawa Laboratory, University of Tokyo and Mori Laboratory, Yokohama National University. This system is based on occurrence and concatenation frequencies of simple and compound nouns (Nakagawa and Mori, 2003). Termex assigns a score that approximates termhood for each extracted technical term and outputs the ranked list of technical terms for an input document. We assume the whole output of termex as candidates of anchor texts without filtering by the scores.

[1] http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb_eng.html

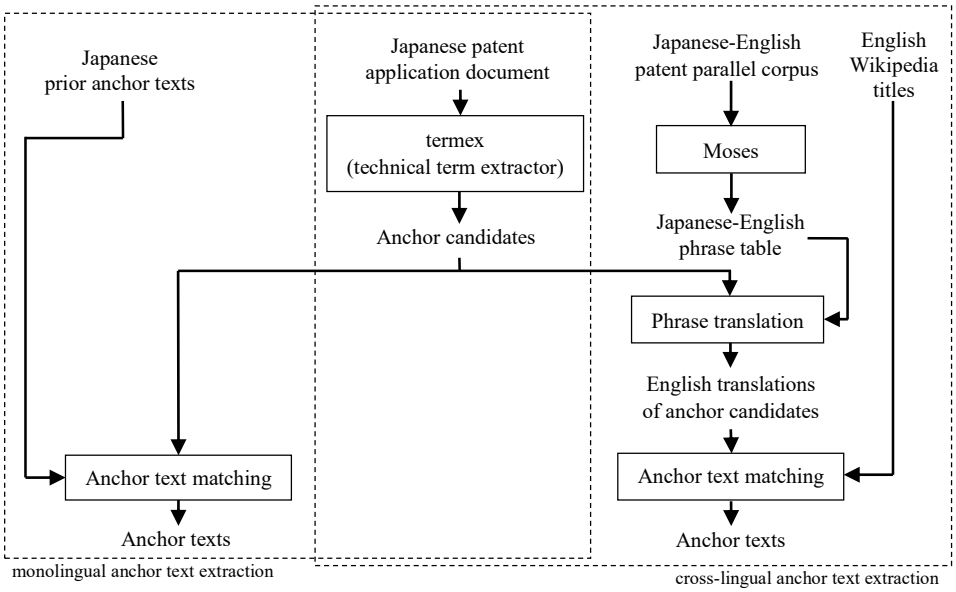

Figure 1. Overview of monolingual and cross-lingual anchor text extraction.

2) Monolingual anchor text extraction

From anchor candidates, we next extract anchor texts by matching with prior anchor texts in the same language. Prior anchor texts are n-grams that have been anchor texts used to link to Wikipedia articles at least once. Each prior anchor text is associated with one or more destination articles; for example, a Japanese prior anchor text "圧縮 (*asshuku*)" has possible destinations "圧縮 (*asshuku*; compression)," "データ圧縮 (*deta asshuku*; data compression)," etc. It is likely that occurrence of the anchor text in a text will be linked to one of the possible destinations.[2] Accordingly, we regard anchor candidates that match one prior anchor text as anchor texts. Extracting more anchor texts indicates the possibility of enriching patent application documents with more links to Wikipedia.

Optionally we use Wikipedia category information to extract only anchor texts related to the domain of the patent application document. Every Wikipedia article belongs to at least one category, which is systematically organized in a taxonomy style. By specifying a category of users' interest, we can only extract anchor texts if one of their destination articles belongs to a category subsumed by the specified category. Another advantage is removing non-technical terms from anchor candidates extracted by the technical term extraction system.

3) Cross-lingual anchor text extraction

In the cross-lingual part, extracted anchor candidates are translated to English by looking up a phrase table constructed from a Japanese-English patent parallel corpus. We generate n-best translations for each anchor candidate by using phrase pairs whose phrase translation probabilities are not below a threshold value θ.[3] If one of the translations matches an English Wikipedia article title, we regard the original Japanese anchor candidate as an anchor text.

[2] Some prior anchor texts may have different meanings from all possible destination articles in current Wikipedia. We ignore such rare cases because they seem not to significantly affect our evaluation results.

[3] We do not apply full phrase-based statistical machine translation to patent application documents because it might cause more translation errors due to the high complexity of sentence-level translation. We consider that phrase-level translation is enough to anchor text extraction for cross-lingual wikification.

The reason we do not use English prior anchor texts is to avoid unexpected matches occurring from translation. For example, consider the Japanese anchor text "細胞 (*saibo*)", which means a cell in biology. The phrase table gives its translation as "cell," and one of the destination articles is "Electrochemical cell," which the Japanese anchor text does not mean. Cross-lingual links may be overestimated when the destination article does not describe the anchor text in a patent application document.

4. Experiment

In this section, we describe an experiment for proving the effectiveness of cross-lingual patent wikification. We compare the numbers of technical terms extracted by monolingual and cross-lingual anchor text extraction. We also confirm that use of Wikipedia categories has effect on disambiguation to Wikipedia articles for anchor texts.

4.1. Data used

In our experiment we used the NTCIR-7 PATMT test collection, which consists of unexamined Japanese patent application documents published in 1993-2002 and USPTO English patent grant data published in 1993-2000. From the test collection, we extracted 10,000 Japanese patent application documents published in 2000 with the international patent classification (IPC) class G06 (COMPUTING; CALCULATING; COUNTING) as the input.

We employed a phrase table provided from a research group in University of Tsukuba, which was used for identification of bilingual synonymous technical terms (Liang et al., 2011). This phrase table was constructed from a Japanese-English patent parallel corpus (Utiyama and Isahara, 2007) in the NTCIR-7 PATMT test collection by using the Moses toolkit (Koehn et al., 2007). Setting the phrase translation probability threshold θ to 0.01, we generated the 10-best translations for each anchor candidate for cross-lingual anchor text extraction.

We used Japanese and English Wikipedia data dumped in March 2013 for collecting Japanese prior anchor texts, English article titles, and category information. For an experiment using categories, we manually specified the category "計算機科学 (Computer science)", which is strongly related to the IPC class cited above. We extract anchor texts if at least one of their destination articles belongs to the categories subsumed by the specified category. To reduce computational complexity, the subsumed categories are limited to categories that can be reached from the specified category by iteratively tracking subcategories within 10 times.

4.2. Anchor text extraction results

Table 1 shows an example of monolingual and cross-lingual anchor text extraction results for anchor candidates extracted by termex from an input patent application document. **Score** indicates a termhood score output by termex. Check marks in the **Mono** and **CL** columns show that the anchor candidate was selected as anchor texts that could be linked to an article by monolingual and cross-lingual anchor text extraction, respectively. For example, the term "復号化 (decoding)" in the second row is checked only in the **CL** column because it was linked to an English article "Decoding" but Japanese Wikipedia did not have an independent article to describe it. The **InCat** column shows whether one of the Japanese destination articles belongs to a category subsumed by the specified category "計算機科学 (Computer science)."

After extracting all anchor texts in the 10,000 patent application documents, we obtained proportions of anchor candidates selected as anchor texts in each interval of the termhood scores: 1-10, 10-100, 100-1000, 1000-3000, 3000-5000, 5000-10000, and 10000-. The total numbers of anchor candidates in each score interval are described in Table 2. Figure 2

Extracted anchor candidate	Score	Anchor text decision		InCat
		Mono	CL	
情報 (information)	2544.40	✓	✓	✓
復号化 (decoding)	2318.54		✓	
遊技プログラム情報 (game program information)	2090.44			
書き込み情報 (write information)	1815.47			
復号手段 (decoder)	1648.58		✓	✓
復号書き込み情報 P (decoding write information P)	1621.06			
復号キー (decoding key)	1464.71		✓	
復号キーK (decoding key K)	1162.16			
復号書き込み情報 (decoding write information)	1128.33			
…				
メモリ (memory)	155.88	✓	✓	✓
自己検証回路 (self verification circuit)	135.04			
遊技 (game)	122.38	✓	✓	
暗号化処理 (encrypting/encoding)	119.73		✓	
キー (key)	119.65	✓	✓	✓
…				
アルゴリズム (algorithm)	39.50	✓	✓	✓
制御 (control)	34.50	✓	✓	
遊技内容 (game content)	33.08			
ＲＯＭライタ (ROM writer)	30.32			

Table 1. Examples of extracted anchor texts.

shows the proportions of anchor candidates selected as anchor texts by monolingual and cross-lingual anchor text extraction. The black bars indicate the proportions calculated only on anchor texts with **InCat** checked, and the whole bars were obtained from all extracted anchor texts. Although termex outputs terms with a score of 1.00 or more, technical terms in the intervals 1-10 and 10-100 are considered to be less important. In interval 10000-, cross-lingual anchor text extraction retrieves 9.7% (60.5% vs. 50.8%) of the technical terms that monolingual extraction does not. In all score intervals, cross-lingual extraction obtained links about 1.3-1.5 times more than monolingual extraction.

With the category specified, similar tendency between proportions and scores was observed. Though the percentage of anchor texts in the subsumed categories strongly depends on the specified category, it ranged from one third to a half. This shows that cross-lingual anchor text extraction is also effective for technical terms with categories limited.

score interval	1-10	10-100	100-1000	1000-3000	3000-5000	5000-10000	10000-
#	1732k	825k	257k	37k	7973	5944	4261

Table 2. Numbers of anchor candidates.

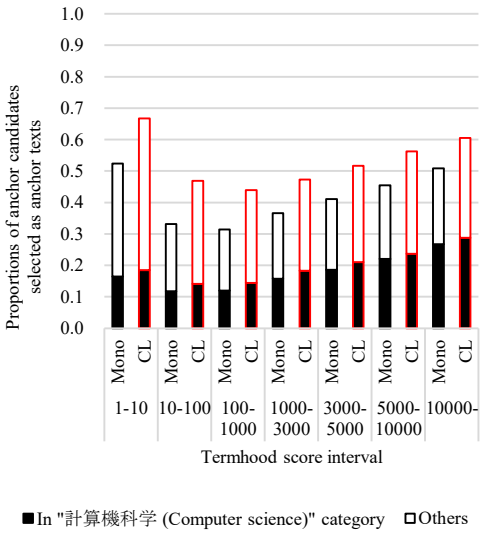

Figure 2. Proportions of anchor candidates selected as anchor texts.

The proportions slightly improve in higher score sections. This indicates that the employed technical term extraction method would be suitable for anchor text extraction of patent wikification, because the system preferentially selects technical terms that are described in Wikipedia, namely, that are of interest to Wikipedians.

4.3. Disambiguation of anchor texts by category information

The wikification process includes a disambiguation step to find a correct destination article for each anchor text. Cross-lingual disambiguation is a more challenging task than monolingual disambiguation because the translation process reduces the impact of most of the effective disambiguation features, such as keyphraseness and context information. Among these features, category information (Cucerzan, 2007) is cross-lingually available because Wikipedia categories have inter-language links across languages: in our case, Japanese category "計算機科学" and English corresponding category "Computer science" are connected by an inter-language link. In our experiment, we simply estimated how category information contributes to disambiguation of anchor texts by eliminating destination articles that does not belong to categories subsumed by the specified category. For example, anchor candidate "キー (key)" in Table 1 has 19 possible destination articles, including "鍵 (key (lock))," "キーボード (コンピュータ) (computer keyboard)," "調 (key (music))," and so on. Among them, only four destination articles such as "キーボード (コンピュータ) (computer keyboard)" belong to the subsumed categories. There were 5.56 possible destination articles on average. After removing destination articles that did not belong to the subsumed categories, only 3.31 destination articles on average remained.

Disambiguation of anchor texts by category information might be strengthened by automatically specifying categories relevant to each patent application document. Using such categories would be more specific than using all categories subsumed by a manually specified category. Hence, it would be able to eliminate more articles irrelevant to an anchor text from the destination candidates than manually specifying a category. One promising method to make this possible is using classification tags for patents, such as tags showing IPC classes. To develop such a method it will be important to find a way to identify Wikipedia categories corresponding to each IPC class.

5. Conclusion and future work

In this paper, we demonstrated that cross-lingual patent wikification is promising for enriching patent application documents by increasing links to Wikipedia articles. The experiment we conducted indicated that more than 60% of important technical terms in patent application documents could be linked to Wikipedia articles with cross-lingual anchor text extraction, while 50% of them with monolingual one.

As a topic for future research, we plan to automatically determine the categories related to each patent application document for improving cross-lingual patent wikification. We also plan to tackle the problem of how to disambiguate anchor texts after applying a category-based method to them.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 15K16096. We express our appreciation to Professor Takehito Utsuro and Professor Mikio Yamamoto at the University of Tsukuba for providing a phrase table built from a patent parallel corpus in the NTCIR data set.

References

Cassidy, Taylor, Heng Ji, Hongbo Deng, Jing Zheng, and Jiawei Han. (2012). Analysis and refinement of cross-lingual entity linking. In *Proceedings of the 3rd International Conference on Information Access Evaluation: Multilinguality, Multimodality, and Visual Analytics*, pages 1-12.

Cucerzan, Silviu. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716.

Jacquemin, Christian and Didier Bourigault. (2003). Term extraction and automatic indexing. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pages 599-615). Oxford University Press.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007) on Interactive Poster and Demonstration Sessions*, pages 177-180.

Liang, Bing, Takehito Utsuro, and Mikio Yamamoto. (2011). Semi-automatic identification of bilingual synonymous technical terms from phrase tables and parallel patent sentences. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 196-205.

McNamee, Paul, James Mayfield, Dawn Lawrie, Douglas W. Oard, and David Doermann. (2011). Cross-language entity linking. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 255-263.

Miao, Qingliang, Huayu Lu, Shu Zhang, and Yao Meng. (2013). Cross-lingual link discovery between Chinese and English wiki knowledge bases. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC)*, pages 374-381.

Mihalcea, Rada and Andras Csomai. (2007). Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM 2007)*, pages 233-242.

Milne, David and Ian H. Witten. (2008). Learning to link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008)*, pages 509–518.

Nakagawa, Hiroshi and Tatsunori Mori. (2003). Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201-209.

Utiyama, Masao and Hitoshi Isahara. (2007). A Japanese-English patent parallel corpus. In *Proceedings of Machine Translation Summit XI*, pages 475-482.