社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007 − DBS − 143 (93)
2007 / 7 / 4

Wikipedia Link Structure Analysis for Extracting Bilingual Terminology

Maike ERDMANN†, Kotaro NAKAYAMA†, Takahiro HARA†, and Shojiro NISHIO†

† Graduate School of Computer Science and Technology, Osaka University
1-5 Yamadaoka, Suita, Osaka, 565-0871, Japan
E-mail: †{erdmann.maike,nakayama.kotaro,hara,nishio}@ist.osaka-u.ac.jp

Abstract With the demand of bilingual dictionaries covering domain-specific terminology, research in the field of automatic dictionary extraction has become popular. However, present research based on the analysis of bilingual text corpora faces several issues regarding very different language pairs such as Japanese and English. Therefore, we present an approach to extracting an English-Japanese dictionary from the link structure of Wikipedia, a huge scale multilingual Web-based encyclopedia. We extracted a large amount of translation pairs and conducted some experiments in which we compared the accuracy and coverage to a dictionary trained on a parallel corpus.

Key words Bilingual Lexicon Extraction, Wikipedia Mining, Web Mining, Link Analysis

1. Introduction

Bilingual dictionaries are required in many research areas, for instance to enhance existing dictionaries with technical terms [9], as seed dictionaries to improve machine translation results, in cross-language information retrieval [3] [8] or for second language teaching and learning. Unfortunately, the manual creation of bilingual dictionaries is not efficient as linguistic knowledge is expensive and new or highly specialized domain specific words are difficult to cover.

In recent years, a lot of research has been conducted on the automatic extraction of bilingual dictionaries. Especially the analysis of large amounts of bilingual text corpora is an emerging research area. However, that approach faces several issues. Particularly, for very different languages or for domains where sufficiently large text corpora are not available, accuracy and coverage of translation dictionaries are rather low.

Therefore, in order to provide a high accuracy and high coverage dictionary, we propose the extraction of bilingual terminology from multilingual encyclopedias like Wikipedia. Wikipedia is a very promising resource as the continuously growing encyclopedia already contains more than 5 million articles in several hundred languages and a broad variety of topics. We already proved that Wikipedia can be used to create an accurate association thesaurus [6] since it has a very dense link structure.

In addition, Wikipedia has a lot of links between articles in different languages. If we regard the titles of Wikipedia articles as terminology, it is easy to extract translation relations by analyzing the interlanguage links, assuming that two articles connected by an interlanguage link are likely to have the same content and thus an equivalent title.

On the other hand, an article in the source language has usually at most one interlanguage link to an article in the target language. Thus, creating a dictionary from interlanguage links only leads to a low coverage for cases where several correct translations for a term exist.

Therefore, we propose two methods to improve the coverage while maintaining a high accuracy. The first method uses redirect pages and the second method uses link texts to extend the number of translations for a given term. In order to evaluate these methods, we extracted Japanese translations for 147 English sample terms and compared the accuracy and coverage of these translations to the translations extracted from a parallel corpus.

The remainder of this paper is organized as follows. We will give an overview on manual dictionary construction and on the state of art in automatic dictionary construction from bilingual texts in section 2 and present our approach in section 3. Then, in section 4, we will describe the experiment we conducted to evaluate our methods and discuss its results. Finally, we will conclude the paper in section 5.

2. Related Work

For bilingual dictionary construction, we can distinguish two approaches: manual and automatic dictionary creation. We will discuss both approaches in the following subsections.

2.1 Manual Dictionary Construction

The traditional way of creating bilingual dictionaries is the manual compilation by human effort. Nowadays, paper-based dictionaries are being more and more replaced by machine readable dictionaries. Besides, those dictionaries are often not created by linguists but voluntarily by a large community of second language learners and other users.

For translations from English to Japanese, one of the most commonly used dictionaries is the freely available online dictionary EDICT. The JMdict/EDICT project [1] was started in 1991 by Jim Breen and the dictionary file has been extended by a large amount of people since then. It comprises more than 99,300 terms as of 2004 including even an impressive large amount of entries for domain-specific terms.

However, even with the aid of a large community, the manual creation of a dictionary is a time-consuming process. In the case of EDICT, it took over 10 years and the effort of numerous people to achieve the current dictionary size. Even though it now covers an impressively high number of terms, latest terms and domain-specific terms are not covered exhaustively. In addition, the correctness of dictionary entries is not guaranteed when e.g. language learners participate, thus the refinement of dictionary entries is time-consuming as well.

2.2 Automatic Dictionary Construction

Nowadays, a lot of machine readable documents in multiple languages are being created every day and often published on the Internet for everyone to access. That has lead to the idea of automatically creating bilingual dictionaries using these resources, thus reducing the burden of manual dictionary compilation. In todays research, mainly two approaches can be distinguished. The first approach uses parallel corpora, bilingual text collections consisting of texts in one language and their translations into another language. For Japanese-English dictionary extraction, e.g. corpora of paper abstracts [10] [3] have been exploited.

However, while for high frequency terms usually good results can be achieved, the accuracy decreases drastically when the term to be translated is not present in the corpus in a large quantity. This is often the case for domain-specific terms.

Furthermore, the accuracy of these dictionaries is rather low for language pairs from very different language families like Japanese and English, since the construction relies on natural language processing. Fung and McKeown [4] stated that for instance, in Asian languages sentence boundaries tend to be in different places than in sentences of European languages. Besides, a parallel corpus does not contain exact translations. For grammatical reasons, or just in order to add supplementary information not generally known by the readers of one language version, some text can be added. Respectively, text can be omitted or presented in a different way in one language.

Another problem is that not for all domains and all languages sufficiently large parallel corpora are available, thus the coverage of the dictionary remains insufficient. Also the collection, e.g. due to copyright restrictions, preparation and analysis of large parallel corpora can be troublesome.

For these reasons, for languages pairs such as Japanese and English, the use of comparable corpora is also interesting. A comparable corpus contains not exact translations but texts from the same domain. Thus we can assume that similar terminology is covered. Among others, research using a corpus of Japanese patent abstracts and their English translations [9] or research using newspaper articles [8] [5] have been conducted. Although it is much easier to collect a comparable corpus than a parallel corpus, it is even more difficult to obtain a sufficient accuracy.

Altogether, the usage of parallel or comparable corpora for automatic dictionary construction is a very interesting approach. However, achieving a sufficient accuracy and coverage is still difficult for less frequent terms as well as for certain language pairs and text domains.

3. Proposed Method

Our idea is to use a multilingual Web-based encyclopedia such as Wikipedia for extracting bilingual terminology. Wikipedia currently contains more than 5 million articles. It covers general topics, domain specific topics as well as proper nouns, containing even latest terminology since Wikipedia is being updated all the time. Moreover, Wikipedia contains a lot of links among its articles, not only within the articles of one language but also between articles of different languages. As opposed to the plain text in bilingual corpora, Wikipedia links contain to some extend semantic information. For instance, an interlanguage link indicates that one page title is the translation of the other. This can decrease difficulties of dictionary creation caused by natural language processing issues.

Wikipedia is being created manually by a large number of contributors. However, we can reuse the contributions for the creation and maintenance of the translation dictionary, thus no additional human effort is needed. Apart from that, when using a pivot language such as English, we can translate words from and to minor languages even if there is no direct translation.

3.1 Wikipedia Link Structure

In order to create a high accuracy and high coverage dictionary, we analyzed several kinds of link information. Prior to describing our methods, we will illustrate the used link structure information in the following clauses.

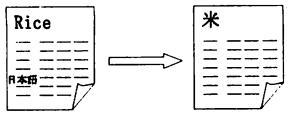

Figure 1 Interlanguage Link

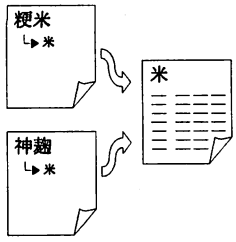

Figure 2 Redirect Pages

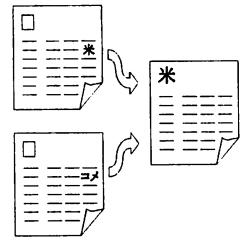

Figure 3 Different Link Texts

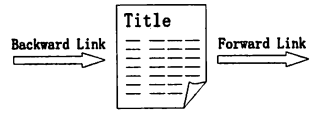

Figure 4 Forward and Backward Links

3.1.1 Interlanguage Links

An interlanguage link in Wikipedia is a link between two articles of the same content in different languages as shown in Figure 1.

Interlanguage links are usually displayed in the left sidebar of an article and are created by using the syntax `[[language code:Title]]`. The language code identifies the language in which the target article is written. The title is the target pages's title. Since the titles of all Wikipedia articles in one language are unique, that information is sufficient to identify the target page unambiguously.

We assume that in most cases, the titles of two articles connected by an interlanguage links are translations of each other.

3.1.2 Redirect Pages

Redirect pages in Wikipedia, shown in Figure 2, are pages containing no content but a link to another article (target page) in order to facilitate the access to Wikipedia content.

When a user accesses a redirect page, he will automatically be redirected to the target page. A redirect page can be created by writing the text `#REDIRECT [[pagename]]` at the top of the article. Pagename contains the name of the target page.

Redirect pages are usually strongly related to the concept of the target page. They can indicate synonym terms, but also abbreviations, more scientific or more comment terms, frequent misspellings or alternative spellings etc.

3.1.3 Link Texts

A link text, shown in Figure 3, is the text part of a link, i.e. the text that is presented to the user in the browser and where he clicks on to reach the target page.

In Wikipedia, when using the default syntax `[[pagename]]`, the title of the target article is displayed as link text. However, link texts can be changed freely by creating a piped link expressed by the syntax `[[pagename|link text]]`.

We extract the link text information by analyzing all internal links (that are links within one language version of Wikipedia) to extract link text information. We already realized that link texts can be used for synonym term extraction [6]. Link texts are usually strongly related to the target page title. In many cases, they differ only in capitalization, but sometimes they are changed in other ways to fit in the sentence structure of the linking article. Therefore, they can help to overcome NLP problems such as finding a translation for a term in plural form when there is only a dictionary entry for the singular form. In some cases however, link text contains unprofitable terms, such as style information in form of HTML tags.

3.1.4 Forward/Backward Links

For all the above mentioned kinds of links, we distinguish the link direction. As shown in Figure 4, a forward link is an outgoing link and a backward link is an incoming link of an article. Both forward and backward links are useful information for extracting translation candidates. Furthermore, the number of backward links is a valuable factor for estimating the probability of a translation candidate as we will describe in the following subsections.

3.2 Extraction of Translation Candidates

At first, we create a baseline dictionary from Wikipedia by extracting all translation candidates from interlanguage links.

In order to do that, for any given term s, a corresponding Wikipedia source page sp is extracted if s is equivalent to the title of that page. In cases where s is equivalent to the title of a redirect page, the corresponding target page is used as sp.

In the second step, in case sp has an interlanguage link to a page tp in the target language, the title t of that page will be used as the translation.

In the following clauses, we will describe two methods for enhancing the baseline dictionary: the RP (redirect page) method and the LT (link text) method.

3.2.1 Enhancement by Redirect Pages

The idea of the RP method is to enhance the dictionary with the redirect page titles R for a set of all redirect pages RP of page tp. The list of translation candidates TC is hence defined as:

$$TC = \{t\} \cup R \, .$$

As mentioned before, not all redirect pages are suitable translations. Therefore, we want to assign a probability value to tp and all extracted redirect pages and filter doubtful terms through a threshold.

We assume that the number of backward links of a page can be used to estimate the accuracy of a translation candidate, because redirect pages where the title is wrong or semantically not related to the title of the target page usually have a small number of backward links. This approach has already proved effective in creating the Wikipedia Thesaurus [6].

We calculate the probability of a redirect page title rp by comparing the number of backward links of rp to the sum of backward links of tp and all its redirect pages. The probability value p is hence defined by the formula:

$$p = \frac{|\text{Backward links of } rp|}{\sum_{rp_i \in \{t\} \cup RP} (|\text{Backward links of } rp_i|)} \, .$$

We can calculate the probability of the target page tp in an analogous manner. Usually, redirect pages have much less backward links than target pages. However, redirect pages with more backward links than the corresponding target pages also exist, indicating that the redirect page title is a good translation candidate, potentially even better than the target page title.

3.2.2 Enhancement by Link Texts

The LT method enhances the dictionary created from interlanguage links with the set of link texts LT of all backward links of tp. The list of translation candidates is thus defined as:

$$TC = \{t\} \cup LT \, .$$

Like for the RP method, we filter unsuitable translations extracted by the LT method by setting a threshold. We calculate the probability of a link text lt by comparing the number of backward links of tp containing the link text lt to the total number of backward links of tp:

$$p = \frac{|\text{Backward links of } tp \text{ with link text } lt|}{|\text{Backward links of } tp|} \, .$$

4. Evaluation

We conducted an experiment in which we compared the translations of 147 terms extracted by our methods to the translations extracted from a parallel *corpus*. In the following, we will describe the experiment and discuss its results.

4.1 Extraction from Wikipedia

We downloaded the English and Japanese Wikipedia database dump data from November/December 2006 [13] containing 3,068,118 English and 455,524 Japanese articles. From that data, we extracted all interlanguage links, link texts and redirect links as well as the number of backward links for each page. In total, we extracted 1,345,318 English and 91,898 Japanese redirect pages, 7,215,301 different English and 2,019,874 different Japanese link texts. In order to improve the accuracy, we applied several thresholds to filter terms with a low probability.

4.2 Extraction from a Parallel Corpus

We compared the translations extracted by our approach to a dictionary extracted from the parallel corpus JENAAD [11]. With 150,000 one-to-one sentence alignments sentences in each language, that corpus consisting of Japanese and English versions of Yomiuri newspaper articles is relatively large compared with other Japanese-English parallel corpora. The corpus has the advantage of being already sentence-aligned (each sentence in one language is paired with exactly one equivalent sentence in the other language) and the Japanese text is split into chunks, a procedure that is indispensable to isolate terms since the Japanese language does not use word boundaries.

We trained the parallel corpus on the open source training tool GIZA++ [7], which is using the IBM Models 1-5 [2] and the Hidden Markov Model [12], both standard models in word alignment research.

The translation candidates were then extracted from the inverse probability table created by GIZA++. Each line of the table consists of a word in the source language, a translation and a probability value. In total, we extracted 1,033,086 translation pairs. The coverage of the dictionary, however, is much smaller than expected from the number of translation pairs, since it contains a lot of noise, i.e wrong translations with very low probability values. In order to improve the accuracy, it was therefore crucial to define thresholds to filter terms with a low probability.

4.3 Term Categories

The experiment was conducted on 147 English terms, exclusively consisting of nouns since the titles of Wikipedia articles usually are nouns. Apart from that, only terms consisting of one word were selected because the dictionary created by GIZA++ does not translate word compounds. The terms were divided into two categories.

67 terms were high frequency terms which we selected semi-automatically using the most frequent nouns in the parallel corpus.

80 terms were low frequency terms. These terms were chosen by native speakers and people fluent in English. These persons were asked to list up technical terms found in English newspapers. We call these terms low frequency terms since they appear in the parallel corpus much less frequent than the terms in the first category, even though the term selectors were not instructed to choose low frequency words. We further split the low frequency terms into two categories with 34 terms that can be found in the dictionary EDICT and 46 terms that cannot be found in EDICT.

4.4 Comparison Criteria

We used the two standard criteria precision and recall to compare accuracy and coverage of our methods and the parallel corpus approach.

The precision measures the accuracy by calculating how many of the extracted translation candidates are correct:

$$precision = \frac{|\text{Correct translations}|}{|\text{All Translations}|} .$$

The recall measures the coverage by calculating how many correct translations have been extracted by a method compared to the total number of correct translations. It is not trivial to estimate the total number of correct translations, since it cannot be calculated automatically. In our experiment, we estimated the value by using the manually counted number of correct translations from EDICT, since it contains a large amount of translations not only for general but also for domain-specific terms:

$$recall = \frac{|\text{Correct translations}|}{|\text{Correct translations in EDICT}|} .$$

The term evaluation as well as the counting of correct EDICT translations were conducted by totally 12 judges, mostly native speakers of Japanese with a sufficient English proficiency.

4.5 Experiment Results

In the following, we will discuss the results of our experiment based on the precision and recall values as well as the absolute number of extracted correct translations.

4.5.1 High Frequency Terms

For high frequency terms, as shown in Table 1, our methods are only slightly better than the parallel corpus approach.

Using only interlanguage links leads to roughly the same result as the parallel corpus approach. Using the RP method, the recall increases only slightly and in exchange the precision decreases a little. For the LT method, the recall is even better than that of the RP approach, but the precision becomes very low. By using thresholds, for both methods precision and recall converge to the values of using interlanguage links only.

Table 1 Precision and Recall for High Frequency Terms

Method	Precision	Recall
Interlanguage Links only	0.846	0.113 (44)
Interlanguage Links With RP (All)	0.635	0.164 (64)
Interlanguage Links With RP ($p > 0.001$)	0.682	0.154 (60)
Interlanguage Links With RP ($p > 0.1$)	0.811	0.11 (43)
Interlanguage Links With LT (All)	0.379	0.272 (106)
Interlanguage Links With LT ($p > 0.001$)	0.519	0.208 (81)
Interlanguage Links With LT ($p > 0.1$)	0.852	0.118 (46)
Parallel Corpus (Top 1)	0.716	0.123 (48)
Parallel Corpus ($p > 0.5$)	0.846	0.085 (33)
Parallel Corpus ($p > 0.1$)	0.655	0.2 (78)
Parallel Corpus ($p > 0.01$)	0.315	0.377 (147)

Table 2 Precision and Recall for Low Frequency Terms in EDICT

Method	Precision	Recall
Interlanguage Links only	0.789	0.185 (15)
Interlanguage Links With RP (All)	0.37	0.21 (17)
Interlanguage Links With RP ($p > 0.001$)	0.531	0.21 (17)
Interlanguage Links With RP ($p > 0.1$)	0.778	0.173 (14)
Interlanguage Links With LT (All)	0.33	0.37 (30)
Interlanguage Links With LT ($p > 0.001$)	0.429	0.333 (27)
Interlanguage Links With LT ($p > 0.1$)	0.8	0.198 (16)
Parallel Corpus (Top 1)	0.37	0.123 (10)
Parallel Corpus ($p > 0.5$)	0.444	0.099 (8)
Parallel Corpus ($p > 0.1$)	0.241	0.16 (13)
Parallel Corpus ($p > 0.01$)	0.103	0.198 (16)

4.5.2 Low Frequency Terms in EDICT

For low frequency terms which can be found in EDICT, the advantage of our approach becomes more apparent as shown in Table 2.

If we use interlanguage links only, the precision is much higher than for the parallel corpus approach while the recall is not much different. Adding redirect pages does not change the recall notably. For the LT method however, a higher recall and a lower precision is obtained if no threshold is given.

4.5.3 Low Frequency Terms not in EDICT

Table 3 shows that our approach is especially effective for low frequency terms which cannot be found in EDICT. Since we cannot calculate the recall value for this category, we only use the absolute number of extracted terms to measure the coverage.

Using interlanguage links only, both coverage and precision are much better than for the parallel corpus approach. Additionally, both RP and LT method lead to an increase in coverage with only a slight change in the precision value.

5. Conclusion and Future Work

In this paper, we presented our approach of bilingual terminology extraction from Wikipedia. We proposed two

Table 3 Precision and Recall for Low Frequency Terms not in EDICT

Method	Precision	Recall
Interlanguage Links only	0.667	- (18)
Interlanguage Links With RP (All)	0.523	- (34)
Interlanguage Links With RP ($p > 0.001$)	0.593	- (32)
Interlanguage Links With RP ($p > 0.1$)	0.676	- (23)
Interlanguage Links With LT (All)	0.563	- (63)
Interlanguage Links With LT ($p > 0.001$)	0.551	- (54)
Interlanguage Links With LT ($p > 0.1$)	0.697	- (23)
Parallel Corpus (Top 1)	0.462	- (6)
Parallel Corpus ($p > 0.5$)	0.6	- (6)
Parallel Corpus ($p > 0.1$)	0.292	- (7)
Parallel Corpus ($p > 0.01$)	0.117	- (7)

methods for extracting terminology. The RP method combines interlanguage link with redirect page information whereas the LT method uses interlanguage links in combination with link text information.

Our methods are very useful for specialized domain-specific terms because our coverage is much better than that of the parallel corpus approach and even better than that of EDICT. Additionally, we may be able to obtain even better results by combining the RP and LT approach, since a term which is both a redirect page title and a link text seems to be a promising translation candidate. Another important aspect is that we evaluated the methods only for single words. Domain-specific words are often word compounds and for those both accuracy and coverage are probably much better than for the parallel corpus approach.

Apart from that, we believe that the encyclopedia will become even much more comprehensive in near future which will also result in a better coverage.

For general terms especially for high frequency terms or for word groups other than nouns, we can probably get good results by combining our approach with the parallel corpus approach or by enhancing it with EDICT entries. Another possiblity is to integrate other, more specialized Wiki-based encyclopedias.

We are planning to further enhance the accuracy and coverage of our translation dictionary by analyzing the redirect pages and link texts of the source language. It is also promising to find ways to extract translation candidates even when the interlanguage links are missing.

We are also planning to extract a complete English-Japanese dictionary and possibly dictionaries for other language pairs.

Acknowledgments

This research was supported in part by Grant-in-Aid on Priority Areas (18049050), and by the Microsoft IJARC CORE project.

References

[1] Breen, J. W.: JMdict: a Japanese-Multilingual Dictionary., *COLING Multilingual Linguistic Resources Workshop* (2004).

[2] Brown, P. F., Pietra, V. J. D., Pietra, S. A. D. and Mercer, R. L.: The mathematics of statistical machine translation: parameter estimation, *Proceedings of the International Conference on Computational Linguistics*, Vol. 19, No. 2, pp. 263–311 (1993).

[3] Chen, Y., Wu, E. and Sun, J.: Automatic Meshing Scheme for Radiosity Calculation of Large-Scale Application, *Ruan Jian Xue Bao (Journal of Software)*, Vol. 10, No. 4, pp. 449–454 (1999).

[4] Fung, P. and McKeown, K.: Aligning Noisy Parallel Corpora Across Language Groups : Word Pair Feature Matching by Dynamic Time Warping, *eprint arXiv:cmp-lg/9409011* (1994).

[5] Kaji, H.: Adapted seed lexicon and combined bidirectional similarity measures for translation equivalent extraction from comparable corpora, *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 115–124 (2004).

[6] Nakayama, K., Hara, T. and Nishio, S.: A Thesaurus Construction Method from Large ScaleWeb Dictionaries., *International Conference on Advanced Information Networking and Applications*, pp. 932–939 (2007).

[7] Och, F. J. and Ney, H.: Improved Statistical Alignment Models, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 440–447 (2000).

[8] Sadat, F., Yoshikawa, M. and et al.: Bilingual Terminology Acquisition from Comparable Corpora and Phrasal Translation to Cross-Language Information Retrieval, *The Companion Volume to the Proceedings of Annual Meeting of the Association for Computational Linguistics*, pp. 141–144 (2003).

[9] Shimohata, S.: Finding Translation Candidates from Patent Corpus, *Proceedings of the Machine Translation Summit*, pp. 50–54 (2005).

[10] Tsuji, K. and Kageura, K.: Automatic generation of Japanese-English bilingual thesauri based on bilingual corpora, *Journal of the American Society for Information Science and Technology*, Vol. 57, No. 7, pp. 891–906 (2006).

[11] Utiyama, M. and Isahara, H.: Reliable Measures for Aligning Japanese-English News Articles and Sentences., *Proceedings of the Annual Meeting of Association for Computational Linguistics*, pp. 72–79 (2003).

[12] Vogel, S., Ney, H. and Tillmann, C.: HMM-based word alignment in statistical translation, *Proceedings of the Conference on Computational Linguistics*, pp. 836–841 (1996).

[13] Wikimedia Foundation: Wikimedia Downloads, *http://download.wikimedia.org/*.