IPSJ Journal Vol. 49 No. 7 2468–2479 (July 2008)

Regular Paper

Extraction of Bilingual Terminology from a Multilingual Web-based Encyclopedia

Maike Erdmann,[†1] Kotaro Nakayama,[†1]
Takahiro Hara[†1] and Shojiro Nishio[†1]

With the demand for bilingual dictionaries covering domain-specific terminology, research in the field of automatic dictionary extraction has become popular. However, the accuracy and coverage of dictionaries created based on bilingual text corpora are often not sufficient for domain-specific terms. Therefore, we present an approach for extracting bilingual dictionaries from the link structure of Wikipedia, a huge scale encyclopedia that contains a vast number of links between articles in different languages. Our methods analyze not only these interlanguage links but extract even more translations from redirect page and link text information. In an experiment which we have interpreted in detail, we proved that the combination of redirect page and link text information achieves much better results than the traditional approach of extracting bilingual terminology from parallel corpora.

1. Introduction

Bilingual dictionaries are required in many research areas, for instance to enhance existing dictionaries with technical terms[16], as seed dictionaries to improve machine translation results, in cross-language information retrieval[15] or for second language teaching and learning. Unfortunately, the manual creation of bilingual dictionaries is inefficient since 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. In particular 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 such as Wikipedia. Wikipedia is a very promising resource as the continuously growing encyclopedia already contains more than 10 million articles in more than 200 languages and covers a wide variety of topics. We have already proved that Wikipedia can be used to create an accurate association thesaurus[12],[13] because of its 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 equivalent titles.

Interlanguage links have already been used to extract bilingual terminology[1],[4]. However, 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 new methods to improve the coverage while maintaining a high accuracy. Our methods use redirect pages and link texts to extend the number of translations for a given term. In order to evaluate our methods, we extracted Japanese translations for 200 English sample terms and compared accuracy and coverage of these translations with the translations extracted from a parallel corpus.

The paper is organized as follows. We 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. In Section 4, we describe the experiment we conducted to evaluate our methods and discuss its results. Finally, we conclude the paper in Section 5.

2. Related Work

For bilingual dictionary construction, we can distinguish two approaches: man-

[†1] Department of Multimedia Engineering, Graduate School of Information Science and Technology, Osaka University

ual and automatic dictionary construction. We 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 replaced more and more 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[2] was started in 1991 by Jim Breen and the dictionary file has been extended by a large number of people since then. It comprises more than 99,300 terms as of 2004 including an impressive large number 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.

A lot of research has been conducted on the extraction of bilingual terminology from parallel corpora, bilingual text collections consisting of texts in one language and their translations into another language. Very promising results have been achieved with the IBM models[3] as well as the Hidden Markow Model[19], which were originally developed for machine translation but can also be used to translate single terms. Another notable approach by Melamed[10),11] is much simpler and tailored for the translation of single words. He achieves an impressively high accuracy but the assumption of his method that each term has only one correct translation naturally leads to a low recall.

One of the main issues of bilingual dictionary extraction from parallel corpora is that while good results for high frequency terms can usually be achieved, the accuracy decreases drastically when the term to be translated is not often present in the corpus. 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. For instance, in Asian languages sentence boundaries tend to be in different places than in sentences of European languages. Besides, Fung and McKeown[5] stated that a parallel corpus often 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 in dictionary extraction from bilingual corpora is that sufficiently large parallel corpora are not sufficiently available for all domains and all languages, 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 Japanese-English dictionary extraction, e.g., corpora of paper abstracts[17] or software documentations[5] have been used. However, since the number of Japanese-English parallel corpora is very limited, 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 with non-verbatim English translations[16] and research using newspaper articles[7),15] has been conducted. Although it is much easier to collect a comparable corpus than a parallel corpus, it is even more difficult to obtain sufficient accuracy.

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

3. Proposed Methods

Our idea is to use a multilingual Web-based encyclopedia such as Wikipedia for extracting bilingual terminology. Wikipedia currently contains more than 10 million articles. It covers general topics, domain-specific topics as well as named entities, containing even latest terminology since Wikipedia is being updated all the time. Moreover, Wikipedia contains many 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 extent 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, and thus no additional human effort is needed.

3.1 Wikipedia Link Structure

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

3.1.1 Interlanguage Links

An interlanguage link in Wikipedia is a link between two articles in different languages as shown in **Fig. 1**. In most cases, the titles of two articles connected by an interlanguage link are translations of each other.

3.1.2 Redirect Pages

Redirect pages in Wikipedia, shown in **Fig. 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. Redirect pages are usually strongly related to the concept of the target page. They often indicate synonym terms, but can also be e.g., abbreviations, more scientific or more common terms, frequent misspellings or alternative spellings.

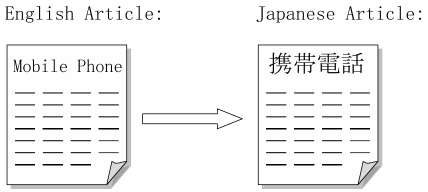

Fig. 1 Interlanguage link example.

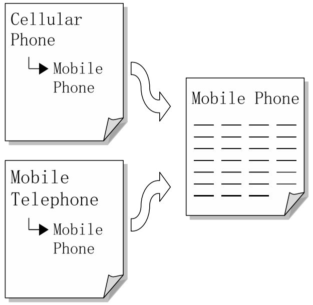

Fig. 2 Redirect page examples.

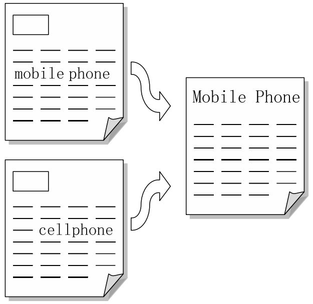

Fig. 3 Link text examples.

3.1.3 Link Texts

A link text, also called anchor text, is the text part of a link that is presented to the user in the browser, as shown in **Fig. 3**, which he clicks on to reach the target page.

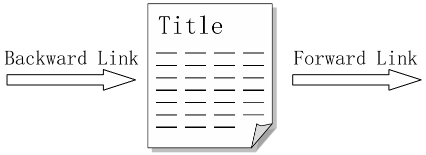

Fig. 4 Forward and backward links.

In Wikipedia, the title of the target article is displayed as the link text by default. However, link texts can be changed freely by creating so called piped links.

We extract the link text information by analyzing all internal links, i.e., links within one language version of Wikipedia. We have already realized that 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 texts contain terms that are not synonyms or include metadata such as HTML tags.

3.1.4 Forward/Backward Links

For all the above mentioned kinds of links, we distinguish the link direction. As shown in **Fig. 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 quality of a translation candidate as we describe in the following subsections.

3.2 Extraction of Translation Candidates

In the following, we describe how we extract a baseline dictionary from interlanguage links and present three methods for enhancing that dictionary; the redirect page method, the link text method, and the combination of both methods. Some of the variables used in the following are visualized in **Fig. 5**.

3.2.1 Extraction of Interlanguage Links

At first, we create a baseline dictionary from Wikipedia by extracting all translation candidates from interlanguage links. The flow is described as follows.

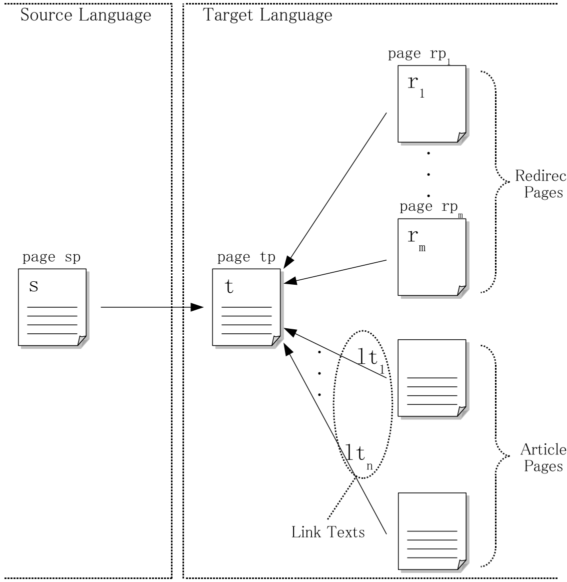

Fig. 5 Wikipedia link structure.

For a term s to be translated, a Wikipedia source page sp is extracted if its title is equivalent to that term. In cases where s is equivalent to the title of a redirect page, the corresponding target page is used as sp. Furthermore, if s is a link text in the source language of Wikipedia, we set the linked page as sp. Thus, for one term to be translated, more than one page in Wikipedia can be utilized as a source page.

In the second step, we try to find interlanguage links for each source page in Wikipedia. In the case where a page sp has an interlanguage link to a page tp in the target language, the title t of tp is chosen as a translation candidate, thus the set of translation candidates TC is defined as:

$$TC(s) = \{t\} .$$

3.2.2 Enhancement by Redirect Pages

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

$$TC(s) = \{t\} \cup R(tp) \ .$$

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

We found out experimentally 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. Recent researches on Web structure mining, such as Google's PageRank [9] and Kleinberg's HITS [8], have already proved the effectiveness of analyzing backward links in order to extract objective and reliable data.

We calculate the score of the redirect page title r of a redirect page rp by comparing the number of backward links of rp to the maximum number of backward links of tp and all its redirect pages.

The score is hence defined by the formula:

$$score_{rp_i} = \frac{\log_e \ bl(rp_i)}{\log_e \ maxbl(tp, rp_1, ..., rp_n)} \ ,$$

where bl calculates the number of backward links for a single page and $maxbl$ calculates the maximum number of backward links for a set of pages.

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

3.2.3 Enhancement by Link Texts

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

$$TC(s) = \{t\} \cup LT(tp) \ .$$

As with the redirect page method, we filter unsuitable translations extracted by the link text method by setting a threshold. We calculate the score of a link text lt by comparing the number of backward links of tp containing the link text lt to the maximum number of backward links of tp containing other link texts:

$$score_{lt_i} = \frac{\log_e \ bl(tp \text{ with } lt_i)}{\log_e \ maxbl(tp \text{ with } lt_1, ..., tp \text{ with } lt_n)} \ .$$

3.2.4 Enhancement by Redirect Pages and Link Texts

In the last method, we combine the redirect page method and the link text method; thus the set of translation candidates TC can be enhanced as follows:

$$TC(s) = \{t\} \cup R(tp) \ \cup LT(tp) \ .$$

The overall score s of a translation candidate $c \in TC$ can now be calculated as follows.

If c is the target page title or a redirect page title and at the same time a link text ($c \in (\{t\} \cup R) \wedge c \in LT$), the score is the weighted sum of $score_{rp}$ and $score_{lt}$:

$$score_c = (w_{rp_c} \cdot score_{rp_c}) + (w_{lt_c} \cdot score_{lt_c}) \ .$$

The variables w_{rp} and w_{lt} represent weight factors to normalize the score. For our experiment, we chose $w_{rp} = w_{lt} = 1$. We also tested unequal weight factors as well as weight factors resulting in larger or smaller sums, but we could not detect a significant influence on the result.

If c is the target page title or a redirect page title but not a link text ($c \in (\{t\} \cup R) \wedge c \notin LT$), the score is calculated by only $score_{rp}$:

$$score_c = w_{rp_c} \cdot score_{rp_c} \ .$$

If c is a link text but neither the target page title nor a redirect page title ($c \in LT \wedge c \notin (\{t\} \cup R)$), the score is calculated by only $score_{lt}$:

$$score_c = w_{lt_c} \cdot score_{lt_c} \ .$$

4. Evaluation

We conducted an experiment in which we compared the translations of 200 terms extracted by our methods to the translations extracted from a parallel corpus. By doing that, we prove that our methods perform better than the traditional and well proven approach of extracting translations from bilingual text corpora. In addition, we also compared the coverage of our methods to EDICT to show that we can extract translations not listed in comprehensive, manually created dictionaries. In the following, we 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[20] containing 3,068,118 English and 455,524 Japanese articles (including redirect pages). From that data, we extracted all interlanguage links, link texts and redirect pages as well as the number of backward links for each page. In total, we extracted 103,374 interlanguage links from English to Japanese, 108,086 interlanguage links from Japanese to English, 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 score.

4.2 Extraction from a Parallel Corpus

We compared the translations extracted by our approach to a dictionary extracted from the parallel corpus JENAAD[18]. We decided to use this corpus since with 150,000 one-to-one sentence alignments 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. Besides, the corpus has the advantage of being already sentence-aligned (each sentence in one language is paired exactly with one 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 used the IBM Models 1-5[3] in combination with the Hidden Markov Model[19] to train the corpus, since these are standard models often used for word alignment.

The training was accomplished using the open source training tool GIZA++[14] and 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 score. 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 candidates, since it contains a lot of noise, i.e., wrong translations with very low scores. In order to improve the accuracy, it was therefore crucial to define thresholds to filter terms with low scores.

4.3 Term Selection

The experiment was conducted on 200 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. 100 terms were high frequency terms which we selected semi-automatically using the most frequent nouns in the parallel corpus. 100 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 frequently 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 50 terms that could be found in the dictionary EDICT and 50 terms that could not be found in EDICT. Example terms are listed in **Table 1**.

4.4 Comparison Criteria

We calculated the two standard criteria precision and recall to compare the 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:

$$precision = \frac{|\text{Extracted correct translations}|}{|\text{All extracted translation candidates}|}\ .$$

The recall measures the coverage by calculating how many correct translations were extracted by a method compared to the total number of correct translations:

$$recall = \frac{|\text{Extracted correct translations}|}{|\text{All correct translations}|}\ .$$

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 using the union of correct translations in EDICT, Wikipedia and the parallel corpus. The calculated recall is therefore a relative recall which is often used in e.g., search engine evaluation[6].

We further evaluated the balance of precision and recall by using the F_α-measure which is defined as:

$$F_\alpha = \frac{(1+\alpha)\cdot(precision \cdot recall)}{\alpha \cdot precision + recall}\ .$$

We calculated the F_1-measure, which weighs precision and recall equally, and

Table 1 Example terms.

High Frequency Terms	Low Frequency Terms in EDICT	Low Frequency Terms not in EDICT
government	sanctions	HIV
year	shareholder	bipartisanship
system	immigration	halftime
people	mayor	Republican
law	tuberculosis	populist
military	communism	Balkan
help	franchise	forensics
money	legislation	EU
industry	lobbyist	rogue
market	amnesty	filibuster
. . .	. . .	. . .

Table 2 Example translations extracted from Wikipedia.

Translation Quality	Term	Translation	
Correct translation	EU	欧州連合	(European Union)
	chairman	会長	(chairman)
	tuberculosis	結核	(tuberculosis)
Too specific translation	money	小銭	(coins)
Too general translation	grenade	爆弾	(bomb)
Translation with different meaning	Japan	日	(day, abbreviation for Japan)
Correct but unusual translation	child	砂利	(colloquial expression for child)

the $F_{0.1}$-measure, which weighs precision ten times as much as recall.

The term evaluation as well as the counting of correct translations was conducted by 12 judges in total, mostly native speakers of Japanese with sufficient English proficiency.

4.5 Experiment Results

In total 890 different translations for the 200 terms were extracted by our methods. A list of sample translations extracted from interlanguage links, redirect pages and texts can be found in **Table 2**. While many of the extracted translations were correct, some others should not be included in a dictionary.

For instance, some translations had too specific or too general meanings. Such translations were extracted from interlanguage links as well as from redirect pages and link texts. Often, the reason is that in cases where a required article does not exist, sometimes an article with more general or more specific content is linked.

Other translations had even more different meanings. These were often extracted from link texts, since link texts have to be adapted to fit into the context of the sentence.

A small number of translations were correct in principle but rarely used, such as antiquated or colloquial expressions. They were sometimes extracted from redirect pages, since redirect pages are intended to facilitate access to Wikipedia content by forwarding all article requests to the actual article.

Incorrect translations were extracted from redirect page and link text information much more often than from interlanguage links, for which reason it proved effective to filter the translation candidates by thresholds.

In the following, we discuss the results of our experiment based on precision and recall, shown in **Figs. 6** and **7**, as well as based on F_1-measure and $F_{0.1}$-measure, shown in **Figs. 8** and **9**. For the parallel corpus approach, the results without using a threshold are not included, because the number of translation candidates would have been too high for manual evaluation.

4.5.1 High Frequency Terms

For high frequency terms, we can see that the combination of redirect page and

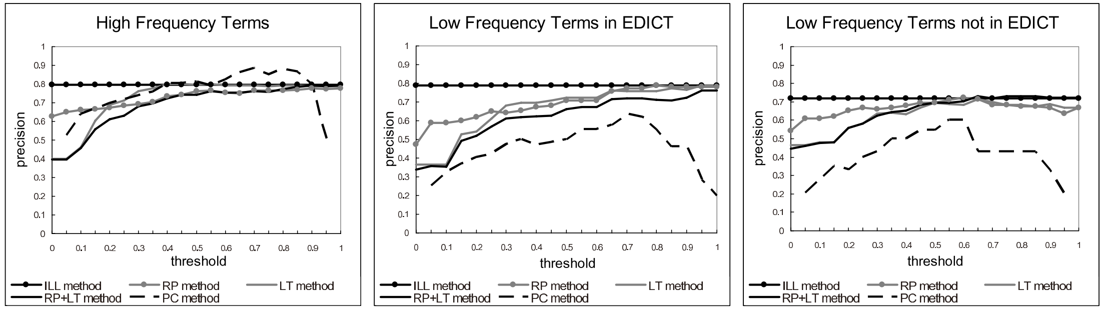

Fig. 6 Precision.

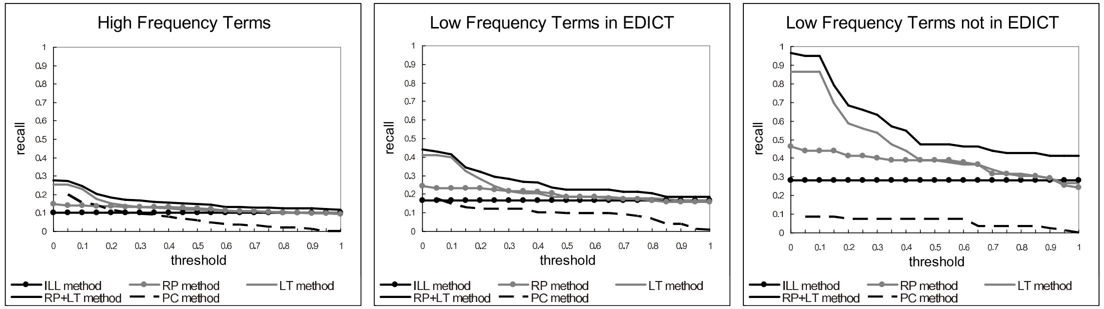

Fig. 7 Relative recall.

link text information (RP ∪ LT method) achieves the highest F_1-measure, thus it is very suitable for applications where the recall is as important as the precision. However, it does not achieve the highest $F_{0.1}$-measure for very low thresholds. Therefore, for applications that value a high precision more than a high recall, the advantage of that method compared to using interlanguage links only (ILL method) is less noticeable.

Compared to the parallel corpus approach (PC method), F_1-measure and $F_{0.1}$-measure of the RP ∪ LT method are both higher; thus it has a clear advantage.

In this category, the overall recall of our methods is very low. Significantly fewer translations than in EDICT are covered. We believe that this is because high frequency terms are often well-known general terms and thus there is no need to cover them in Wikipedia articles. For instance, for the terms "work" and "situation," no translation candidates could be extracted from Wikipedia. Another reason for the low recall is that if the term to be translated is ambiguous such as "party" or "diet," we often cannot find a translation, since disambiguation pages in Wikipedia usually do not contain interlanguage links.

4.5.2 Low Frequency Terms in EDICT

For low frequency terms which are contained in EDICT, the RP ∪ LT method achieves the highest F_1-measure score, thus can be used for applications where both recall and precision are equally important. As for the $F_{0.1}$-measure, our methods do not perform better than the ILL method.

In this category, the performance of our methods is much better than that of

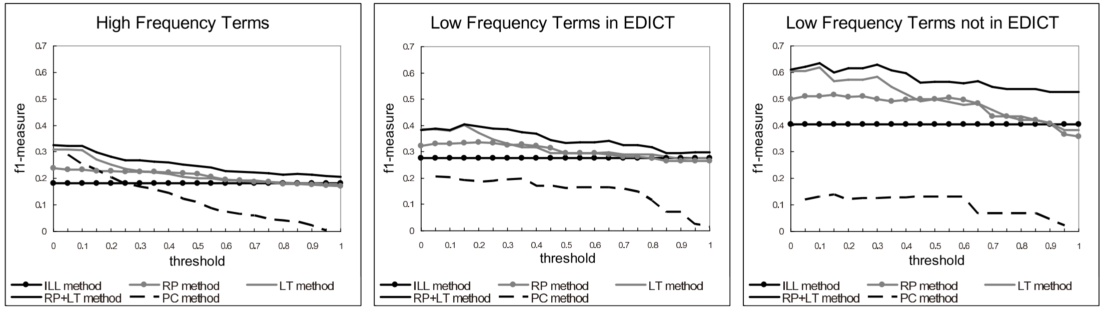

Fig. 8 F_1-measure.

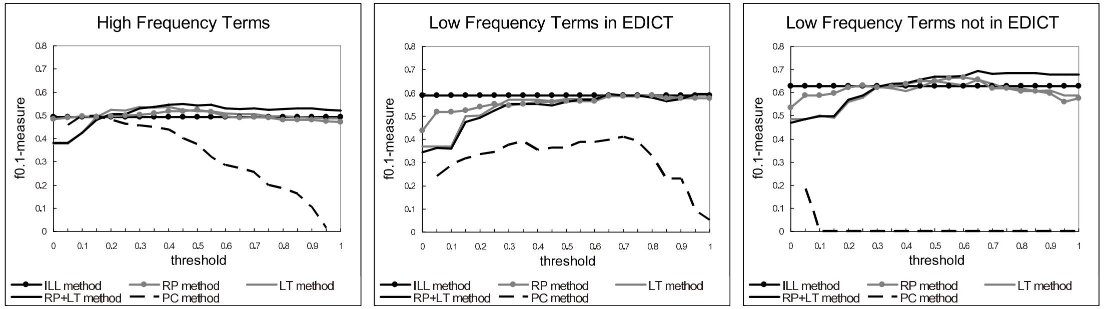

Fig. 9 $F_{0.1}$-measure.

the parallel corpus approach, which can be observed from the F_1-measure and $F_{0.1}$-measure scores.

The recall of our methods is better than for high frequency terms. That is because low frequency terms are often domain-specific terms such as "entrepreneurship" or "communism," whose coverage in Wikipedia is high compared to high frequency terms. For the parallel corpus approach however, good translation results can only be achieved when a term is contained in the corpus in high quantity. For that reason, the parallel corpus approach did not perform very well in our experiment.

4.5.3 Low Frequency Terms not in EDICT

For low frequency terms which could not be found in EDICT, the F_1-measure scores of the RP ∪ LT method are remarkably high compared to those of the ILL method, and also the $F_{0.1}$-measure scores for higher thresholds exceed those of the ILL method.

Furthermore, as in the low frequency terms contained in EDICT, F_1-measure and $F_{0.1}$-measure of the parallel corpus approach are much lower than those of our methods.

The recall in this category is very high, since as explained in Section 4.4, it is not an absolute but a relative recall, and thus used only to compare the different methods with each other.

The results in this term category show that since all terms are not included in EDICT, our methods are also valuable to enhance manually constructed dictionaries. Wikipedia contains very specialized domain-specific terms not covered in EDICT and we thus can extract translations even for terms such as "al-Quaeda" or "OECD."

4.6 Compound Words

In our experiment, we evaluated only translations of single words, although domain-specific terms are often word compounds. However, a number of previously conducted sample translations have shown that there is no reason to be concerned about the accuracy and coverage of our methods for terminology consisting of multiple words. On the contrary, presumably the advantages of our methods compared to bilingual text corpus approaches will become even more significant. In Wikipedia page titles and link texts, we always know which words form a unit. In the plain text of a parallel corpus however, it is rather difficult to determine which words belong together, even when we develop a translation model that can translate word compounds.

5. Conclusion and Future Work

In this paper, we presented our approach of bilingual dictionary extraction from Wikipedia, a multilingual encyclopedia. We proposed three methods for extracting terminology which are using not only interlanguage links but also redirect page and link text information.

Our conviction that Wikipedia is an invaluable resource for bilingual dictionary extraction and that redirect pages and link texts are helpful to enhance a dictionary constructed from interlanguage links has been confirmed in our experiment. Our methods, especially the combination of RP and LT method, have a much better accuracy and coverage than the bilingual text corpus approach. Apart from that, our methods also perform better than the baseline dictionary created from interlanguage links, especially for domain-specific terms and for applications where a high coverage is at least as important as a high accuracy, such as cross-language information retrieval.

Compared to manually created dictionaries, our approach does not perform very well for the translation of general terms. On the other hand, for domain-specific terminology we can extract many accurate translations not covered in manually created dictionaries. In addition, we believe that Wikipedia will become even more comprehensive in the near future which will also result in a better coverage.

It is promising to combine our dictionary with manually constructed dictionaries such as EDICT in order to enhance the coverage for general terms, especially for word groups other than nouns. For applications where not a single term but an entire text has to be translated (e.g., machine translation), we can benefit from combining our approach with the parallel corpus approach.

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.

Our bilingual dictionary can be accessed freely under the URL *http://wikipedia-lab.org*. We are also planning to extract dictionaries for other language pairs.

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

References

1) Bouma, G., Fahmi, I., Mur, J., van Noord, G., van der Plas, L. and Tiedemann, J.: The University of Groningen at QA@CLEF 2006 Using Syntactic Knowledge for QA, *Working Notes for the Cross Language Evaluation Forum Workshop* (2006).

2) Breen, J.W.: JMdict: A Japanese-Multilingual Dictionary, *Proc. COLING Multilingual Linguistic Resources Workshop*, pp.71–78 (2004).

3) Brown, P.F., Pietra, V.J.D., Pietra, S.A.D. and Mercer, R.L.: The Mathematics of Statistical Machine Translation: Parameter Estimation, *Proc. International Conference on Computational Linguistics*, Vol.19, No.2, pp.263–311 (1993).

4) Declerck, T., Pérez, A.G., Vela, O., Gantner, Z. and Manzano-Macho, D.: Multilingual Lexical Semantic Resources for Ontology Translation, *Proc. International Conference on Language Ressources and Evaluation* (*LREC*), pp.1492–1495 (2006).

5) Fung, P. and McKeown, K.: A Technical Word- and Term-Translation Aid Using Noisy Parallel Corpora across Language Groups, *Machine Translation*, Vol.12, No.1-2, pp.53–87 (1997).

6) Goncalves, P., Robin, J., Santos, T., Miranda, O. and Meira, S.: Measuring the Effect of Centroid Size on Web Search Precision and Recall, *Proc. Annual Conference of the Internet Society* (*INET*) (1998).

7) Kaji, H.: Adapted Seed Lexicon and Combined Bidirectional Similarity Measures for Translation Equivalent Extraction from Comparable Corpora, *Proc. Conference on Theoretical and Methodological Issues in Machine Translation*, pp.115–124

(2004).

8) Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment, *J. ACM*, Vol.46, No.5, pp.604–632 (1999).

9) Lawrence, P., Sergey, B., Rajeev, M. and Terry, W.: The PageRank Citation Ranking: Bringing Order to the Web, *Technical Report, Stanford Digital Library Technologies Project* (1999).

10) Melamed, I.D.: A Word-to-Word Model of Translational Equivalence, *Proc. Conference on European Chapter of the Association for Computational Linguistics* (*EACL*), pp.490–497 (1997).

11) Melamed, I.D.: Empirical Methods for MT Lexicon Development, *Proc. Conference of the Association for Machine Translation in the Americas*, pp.18–30 (1998).

12) Nakayama, K., Hara, T. and Nishio, S.: A Thesaurus Construction Method from Large Scale Web Dictionaries, *Proc. IEEE International Conference on Advanced Information Networking and Applications* (*AINA*), pp.932–939 (2007).

13) Nakayama, K., Hara, T. and Nishio, S.: Wikipedia Mining for An Association Web Thesaurus Construction, *Proc. International Conference on Web Information Systems Engineering* (*WISE*), pp.322–334 (2007).

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

15) Sadat, F., Yoshikawa, M. and Uemura, S.: Bilingual Terminology Acquisition from Comparable Corpora and Phrasal Translation to Cross-Language Information Retrieval, *Proc. Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.141–144 (2003).

16) Shimohata, S.: Finding Translation Candidates from Patent Corpus, *Proc. Machine Translation Summit*, pp.50–54 (2005).

17) 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).

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

19) Vogel, S., Ney, H. and Tillmann, C.: HMM-based Word Alignment in Statistical Translation, *Proc. Conference on Computational Linguistics* (*CL*), pp.836–841 (1996).

20) Wikimedia Foundation: Wikimedia Downloads, http://download.wikimedia.org.

(Received October 1, 2007)
(Accepted April 8, 2008)

(Original version of this article can be found in the Journal of Information Processing Vol.16, pp.68–79.)

Maike Erdmann is currently pursuing her Ph.D. in Information Science and Technology at Osaka University, Japan. She received her B.Sc. in Computing Science from CvO University Oldenburg, Germany in 2006, and her Master of Information Science and Technology from Osaka University, Japan in 2008. Her research interests include knowledge extraction from the WWW and natural language processing.

Kotaro Nakayama received his B.I. and M.I. from Kansai University, Japan and his Ph.D. from Osaka University in 2001, 2003 and 2007. While he was an undergraduate student, he launched an IT company named “Kansai Information Institute” and managed the company as the president and one of the directors for three years. During his master course, he was a lecturer of “Internet architecture” at Doshisha Women’s College. After receiving his Ph.D., he became a Ph.D. researcher at Osaka University. Since 2008, he has been an Assistant Professor at the Center for Knowledge Structuring of the University of Tokyo. His research areas are mainly AI and the WWW. He is especially interested in knowledge extraction from huge scale WWW contents. He is a member of ACM, IEEE, JSAI, IPSJ and IEICE.

Takahiro Hara received his B.E., M.E., and D.E. in Information Systems Engineering from Osaka University in Japan, in 1995, 1997, and 2000. He is currently an Associate Professor at the Department of Multimedia Engineering of Osaka University. His research interests include distributed database systems in advanced computer networks, such as high-speed networks and mobile computing environments. Dr. Hara is a member of ACM, IEEE, IEICE, IPSJ, and DBSJ.

Shojiro Nishio received his B.E., M.E., and Dr.E. from Kyoto University, Japan, in 1975, 1977, and 1980. He was with the Department of Applied Mathematics and Physics of Kyoto University from 1980 to 1988. In October 1988, he joined the faculty of the Department of Information and Computer Sciences of Osaka University. He became a full professor at the Department of Information Systems Engineering of Osaka University in August 1992. He has been a full professor at the Department of Multimedia Engineering at the same university since April 2002. He served as the founding director of the Cybermedia Center of Osaka University from April 2000 to August 2003, and served as the dean of the Graduate School of Information Science and Technology of this university from August 2003 to August 2007. He has been serving as a trustee and vice president of Osaka University since August 2007. His current research interests include database systems and multimedia systems. Dr. Nishio has served on the Editorial Board of *IEEE Transactions on Knowledge and Data Engineering* and *ACM Transactions on Internet Technology*, and is currently involved with the editorial board of *Data and Knowledge Engineering*. He is a fellow of IEICE and IPSJ, and he is a member of eight learned societies, including ACM and IEEE.