Bilingual Synonym Identification with Spelling Variations

Takashi Tsunakawa* **Jun'ichi Tsujii*†‡**
*Department of Computer Science,
Graduate School of Information Science and Technology, University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan
†School of Computer Science, University of Manchester
Oxford Road, Manchester, M13 9PL, UK
‡National Centre for Text Mining 131 Princess Street, Manchester, M1 7DN, UK
{tuna, tsujii}@is.s.u-tokyo.ac.jp

Abstract

This paper proposes a method for identifying synonymous relations in a bilingual lexicon, which is a set of translation-equivalent term pairs. We train a classifier for identifying those synonymous relations by using spelling variations as main clues. We compared two approaches: the direct identification of bilingual synonym pairs, and the merger of two monolingual synonyms. We showed that our approach achieves a high pair-wise precision and recall, and outperforms the baseline method.

1 Introduction

Automatically collecting synonyms from language resources is an ongoing task for natural language processing (NLP). Most NLP systems have difficulties in dealing with synonyms, which are different representations that have the same meaning in a language. Information retrieval (IR) could leverage synonyms to improve the coverage of search results (Qiu and Frei, 1993). For example, when we input the query 'transportation in India' into an IR system, the system can expand the query to its synonyms; e.g. 'transport' and 'railway', to find more documents.

This paper proposes a method for the automatic identification of bilingual synonyms in a bilingual lexicon, with spelling variation clues. A bilingual synonym set is a set of translation-equivalent term pairs sharing the same meaning. Although a number of studies have aimed at identifying synonyms, this is the first study that simultaneously finds synonyms in two languages, to our best knowledge.

Let us consider the case where a user enters the Japanese query '*kōjō*' (工場, industrial plant) into a cross-lingual IR system to find English documents. After translating the query into the English translation equivalent, 'plant,' the cross-lingual IR system may expand the query to its English synonyms, e.g. 'factory,' and 'workshop,' and retrieve documents that include the expanded terms. However, the term 'plant' is ambiguous; the system may also expand the query to 'vegetable,' and the system is prevented by the term which is different from our intention. In contrast, the system can easily reject the latter expansion, 'vegetable,' if we are aware of bilingual synonyms, which indicate synonymous relations over bilingual lexicons: (*kōjō*, plant) $\sim$ (*kōjō*, factory) and (*shokubutsu*[1], plant) $\sim$ (*shokubutsu*, vegetable)[2] (See Figure 1). The expression of the translation equivalent, (*kōjō*, plant), helps a cross-lingual IR system to retrieve documents that include the term 'plant,' used in the meaning for *kōjō*, or industrial plants.

We present a supervised machine learning approach for identifying bilingual synonyms. Designing features for bilingual synonyms such as spelling variations and bilingual associations, we train a classifier with a manually annotated bilingual lexicon with synonymous information. In order to evaluate the performance of our method, we carried out experiments to identify bilingual synonyms by two approaches: the direct identification of bilingual synonym pairs, and bilingual synonym pairs merged from two monolingual synonym lists. Experimental results show that our approach achieves the F-scores

[1] *Shokubutsu* (植物) means botanical plant.
[2] '$\sim$' represents the synonymous relation.

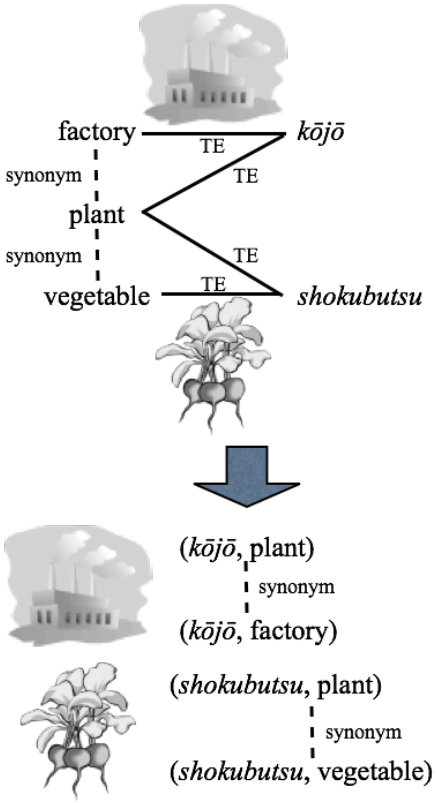

Figure 1: An example of an ambiguous term 'plant', and the synonyms and translation equivalents (TE)

89.3% in the former approach and 91.4% in the latter, thus outperforming the baseline method that employs only bilingual relations as its clues.

The remainder of this paper is organized as follows. The next section describes related work on synonym extraction and spelling variations. Section 3 describes the overview and definition of bilingual synonyms, the proposed method and employed features. In Section 4 we evaluate our method and conclude this paper.

2 Related work

There have been many approaches for detecting synonyms and constructing thesauri. Two main resources for synonym extraction are large text corpora and dictionaries.

Many studies extract synonyms from large monolingual corpora by using context information around target terms (Croach and Yang, 1992; Park and Choi, 1996; Waterman, 1996; Curran, 2004). Some researchers (Hindle, 1990; Grefenstette, 1994; Lin, 1998) classify terms by similarities based on their distributional syntactic patterns. These methods often extract not only synonyms, but also semantically related terms, such as antonyms, hyponyms and coordinate terms such as 'cat' and 'dog.'

Some studies make use of bilingual corpora or dictionaries to find synonyms in a target language (Barzilay and McKeown, 2001; Shimohata and Sumita, 2002; Wu and Zhou, 2003; Lin et al., 2003). Lin et al. (2003) chose a set of synonym candidates for a term by using a bilingual dictionary and computing distributional similarities in the candidate set to extract synonyms. They adopt the bilingual information to exclude non-synonyms (e.g., antonyms and hyponyms) that may be used in the similar contexts. Although they make use of bilingual dictionaries, this study aims at finding bilingual synonyms directly.

In the approaches based on monolingual dictionaries, the similarities of definitions of lexical items are important clues for identifying synonyms (Blondel et al., 2004; Muller et al., 2006). For instance, Blondel et al. (2004) constructed an associated dictionary graph whose vertices are the terms, and whose edges from v_1 to v_2 represent occurrence of v_2 in the definition for v_1. They choose synonyms from the graph by collecting terms pointed to and from the same terms.

Another strategy for finding synonyms is to consider the terms themselves. We divide it into two approaches: rule-based and distance-based.

Rule-based approaches implement rules with language-specific patterns and detect variations by applying rules to terms. Stemming (Lovins, 1968; Porter, 1980) is one of the rule-based approaches, which cuts morphological suffix inflections, and obtains the stems of words. There are other types of variations for phrases; for example, insertion, deletion or substitution of words, and permutation of words such as 'view point' and 'point of view' are such variations (Daille et al., 1996).

Distance-based approaches model the similarity or dissimilarity measure between two terms to find similar terms. The edit distance (Levenshtein, 1966) is the most widely-used measure, based on the minimum number of operations of insertion, deletion, or substitution of characters for transforming one term into another. It can be efficiently calculated by using

Term pairs	Concept
p_1 = (*shōmei* (照明), light)	c_1
p_2 = (*shōmei*, lights)	c_1
p_3 = (*karui* (軽い), light)	c_2
p_4 = (*raito* (ライト), light)	c_1, c_2
p_5 = (*raito*, lights)	c_1
p_6 = (*raito*, right)	c_3
p_7 = (*migi* (右), right)	c_3
p_8 = (*raito*, right fielder)	c_4
p_9 = (*kenri* (権利), right)	c_5
p_{10} = (*kenri*, rights)	c_5

Table 1: An Example of a bilingual lexicon and synonym sets (concepts)

	J terms	E terms	Description
c_1	*shōmei, raito*	light, lights	illumination
c_2	*karui, raito*	light	lightweight
c_3	*migi, raito*	right	right-side
c_4	*raito*	right fielder	(baseball)
c_5	*kenri*	right, rights	privilege

Table 2: The concepts in Table 1

a dynamic programming algorithm, and we can set the costs/weights for each character type.

3 Bilingual Synonyms and Translation Equivalents

This section describes the notion of bilingual synonyms and our method for identifying the synonymous pairs of translation equivalents. We consider a bilingual synonym as a set of translation-equivalent term pairs referring to the same concept.

Tables 1 and 2 are an example of bilingual synonym sets. There are ten Japanese-English translation-equivalent term pairs and five bilingual synonym sets in this example. A Japanese term '*raito*' is the phonetic transcription of both 'light' and 'right,' and it covers four concepts described by the three English terms. Figure 2 illustrates the relationship among these terms. The synonymous relation and the translation equivalence are considered to be similar in that two terms share the meanings. Following synonymous relation between terms in one language, we deal with the synonymous relation between bilingual translation-equivalent term pairs

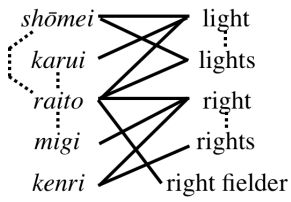

Figure 2: Relations among terms in Table 2
Solid lines show that two terms are translation equivalents, while dotted lines show that two terms are (monolingual) synonyms.

as bilingual synonyms. The advantage of managing the lexicon in the format of bilingual synonyms is that we can facilitate to tie the concepts and the terms.

3.1 Definitions

Let E and F be monolingual lexicons. We first assume that a term $e \in E$ (or $f \in F$) refers to one or more concepts, and define that a term e is a synonym[3] of $e'(\in E)$ if and only if e and e' share an identical concept[4]. Let '$\sim$' represent the synonymous relation, and this relation is not transitive because a term often has several concepts:

$$e \sim e' \ \wedge \ e' \sim e'' \not\Longrightarrow e \sim e''. \tag{1}$$

We define a synonym set (synset) E_c as a set whose elements share an identical concept c: $E_c = \{e \in E | \forall e \text{ refers to } c\}$. For a term set $E_c(\subseteq E)$,

$$\begin{aligned} & E_c \text{ is a synonym set (synset)} \\ \Longrightarrow \quad & \forall e, e' \in E_c \quad e \sim e' \end{aligned} \tag{2}$$

is true, but the converse is not necessarily true, because of the ambiguity of terms. Note that one term can belong to multiple synonym sets from the definition.

Let $D(\subseteq F \times E)$ be a bilingual lexicon defined as a set of term pairs (f, e) $(f \in F, e \in E)$ satisfying that f and e refer to an identical concept. We

[3]For distinguishing from bilingual synonyms, we often call the synonym a monolingual synonym.

[4]The definition of concepts, that is, the criteria of deciding whether two terms are synonymous or not, is beyond the focus of this paper. We do not assume that related terms such as hypernyms, hyponyms and coordinates are kinds of synonyms. In our experiments the criteria depend on manual annotation of synonym IDs in the training data.

call these pairs translation equivalents, which refer to concepts that both f and e refer to. We define that two bilingual lexical items p and $p'(\in D)$ are *bilingual synonyms* if and only if p and p' refer to an identical concept in common with the definition of (monolingual) synonyms. This relation is not transitive again, and if $e \sim e'$ and $f \sim f'$, it is not necessarily true that $p \sim p'$:

$$e \sim e' \ \wedge \ f \sim f' \not\Longrightarrow p \sim p' \tag{3}$$

because of the ambiguity of terms. Similarly, we can define a bilingual synonym set (synset) D_c as a set whose elements share an identical meaning c: $D_c = \{p \in D | \forall p \text{ refers to } c\}$. For a set of translation eqiuvalents D_c,

$$D_c \text{ is a bilingual synonym set (synset)} \Longrightarrow \forall p, p' \in D_c \quad p \sim p' \tag{4}$$

is true, but the converse is not necessarily true.

3.2 Identifying bilingual synonym pairs

In this section, we describe an algorithm to identify bilingual synonym pairs by using spelling variation clues. After identifying the pairs, we can construct bilingual synonym sets by assuming that the converse of the condition (4) is true, and finding sets of bilingual lexical items in which all paired items are bilingual synonyms. We can see this method as the complete-linkage clustering of translation-equivalent term pairs. We can adopt another option to construct them by assuming also that the bilingual synonymous relation has transitivity: $p \sim p' \wedge p' \sim p'' \Longrightarrow p \sim p''$, and this can be seen as simple-linkage clustering. This simplified method ignores the ambiguity of terms, and it may construct a bilingual synonym sets which includes many senses. In spite of the risk, it is effective to find large synonym sets in case the bilingual synonym pairs are not sufficiently detected. In this paper we focus only on identifying bilingual synonym pairs and evaluating the performance of the identification.

We employ a supervised machine learning technique with features related to spelling variations and so on. Figure 3 shows the framework for this method. At first we prepare a bilingual lexicon with synonymous information as training data, and generate a list consisting of all bilingual lexical item

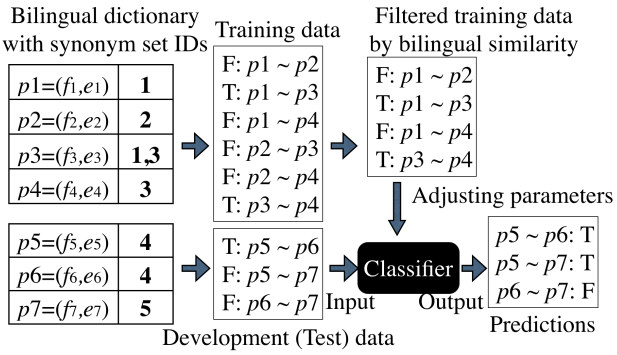

Figure 3: Overview of our framework

pairs in the bilingual lexicon. The presence or absence of bilingual synonymous relations is attached to each element of the list. Then, we build a classifier learned by training data, using a maximum entropy model (Berger et al., 1996) and the features related to spelling variations in Table 3.

We apply some preprocessings for extracting some features. For English, we transform all terms into lower-case, and do not apply any other transformations such as tokenization by symbols. For Japanese, we apply a morphological analyzer JUMAN (Kurohashi et al., 1994) and obtain *hiragana* representations[5] as much as possible[6]. We may require other language-specific preprocessings for applying this method to other languages.

We employed binary or real-valued features described in Table 3. Moreover, we introduce the following combinatorial features: $h_{1F} \wedge h_{1E}$, $\sqrt{h_{2F} \cdot h_{2E}}$, $\sqrt{h_{3F} \cdot h_{3E}}$, $h_{5E} \wedge h_{5F}$, $h_6 \cdot h_{2F}$ and $h_7 \cdot h_{2E}$.

3.2.1 Two approaches for identifying bilingual synonym pairs

There are two approaches for identifying bilingual synonym pairs: one is directly identifying whether two bilingual lexical items are bilingual synonyms ('bilingual' method), and another is first

[5] *Hiragana* is one of normalized representations of Japanese terms, which denotes how to pronounce the term. Japanese vocabulary has many of homonyms, which are semantically different but have the same pronunciation. Despite the risk of classifying homonyms into synonyms, we do not use original forms of Japanese terms because they are typically too short to extract character similarities.

[6] We keep unknown terms of JUMAN unchanged.

h_{1F}, h_{1E}: Agreement of the first characters	Whether the first characters match or not
h_{2F}, h_{2E}: Normalized edit distance	$1 - \frac{\mathrm{ED}(w,w')}{\max(\lvert w\rvert,\lvert w'\rvert)}$, where $\mathrm{ED}(w,w')$ is a non-weighted edit distance between w and w' and $\lvert w\rvert$ is the number of characters in w
h_{3F}, h_{3E}: Bigram similarity	$\frac{\lvert \mathrm{bigram}(w) \cap \mathrm{bigram}(w')\rvert}{\max(\lvert w\rvert,\lvert w'\rvert)-1}$, where $\mathrm{bigram}(w)$ is a multiset of character-based bigrams in w
h_{4F}, h_{4E}: Agreement or known synonymous relation of word sub-sequences	The count that sub-sequences of the target terms match as known terms or are in known synonymous relation
h_{5F}, h_{5E}: Existence of crossing bilingual lexical items	For bilingual lexical items (f_1, e_1) and (f_2, e_2), whether (f_1, e_2) (for h_{5F}) or (f_2, e_1) (for h_{5E}) is in the bilingual lexicon of the training set
h_6: Acronyms	Whether one English term is an acronym for another (Schwartz and Hearst, 2003)
h_7: *Katakana* variants	Whether one Japanese term is a *katakana* variant for another (Masuyama et al., 2004)

Table 3: Features used for identifying bilingual synonym pairs
h_{iF} is the feature value when the terms w and $w'(\in F)$ are compared in the i-th feature and so as h_{iE}. h_6 is only for English and h_7 is only for Japanese.

identifying monolingual synonyms in each language and then merging them according to the bilingual items ('monolingual' method). We implement these two approaches and compare the results. For identifying monolingual synonyms, we use features with bilingual items as follows: For a term pair e_1 and e_2, we obtain all the translation candidates $F_1 = \{f|(f, e_1) \in D\}$ and $F_2 = \{f'|(f', e_2) \in D\}$, and calculate feature values related to F_1 and/or F_2 by obtaining the maximum feature value using F_1 and/or F_2. After that, if all the following four conditions ($p_1 = (f_1, e_1) \in D$, $p_2 = (f_2, e_2) \in D$, $f_1 \sim e_1$ and $f_2 \sim e_2$) are satisfied, we assume that p_1 and p_2 are bilingual synonym pairs[7].

4 Experiment

4.1 Experimental settings

We performed experiments to identify bilingual synonym pairs by using the Japanese-English lexicon with synonymous information[8]. The lexicon consists of translation-equivalent term pairs extracted from titles and abstracts of scientific papers published in Japan. It contains many spelling variations and synonyms for constructing and maintaining the thesaurus of scientific terms and improving the coverage. Table 4 illustrates this lexicon.

	Total	train	dev.	test
$\lvert D\rvert$	210647	168837	20853	20957
$\lvert J\rvert$	136128	108325	13937	13866
$\lvert E\rvert$	115002	91057	11862	12803
Synsets	50710	40568	5071	5071
Pairs	814524	651727	77706	85091

Table 5: Statistics of the bilingual lexicon for our experiment
$|D|, |J|$, and $|E|$ are the number of bilingual lexical items, the number of Japanese vocabularies, and the number of English vocabularies, respectively. 'Synsets' and 'Pairs' are the numbers of synonym sets and synonym pairs, respectively.

Table 5 shows the statistics of the dictionary. We used information only synonym IDs and Japanese and English representations. We extract pairs of bilingual lexical items, and treat them as events for training of the maximum entropy method. The parameters were adjusted so that the performance is the best for the development set. For a monolingual method, we used $T_b = 0.8$, and for a bilingual method, we used $T_b = 0.7$.

4.2 Evaluation

We evaluated the performance of identifying bilingual synonym pairs by the pair-wise precision P,

[7] Actually, these conditions are not sufficient to derive the bilingual synonym pairs described in Section 3.1. We assume this approximation because there seems to be few counter examples in actual lexicons.

[8] This data was edited and provided by Japan Science and Technology Agency (JST).

Synset ID	J term	E term
130213	身体部位 (*shintai-bui*)	Body Regions
130213	身体部位 (*shintai-bui*)	body part
130213	身体部位 (*shintai-bui*)	body region
130213	身体部分 (*shintai-bubun*)	body part
130217	Douglas 窩 (Douglas-*ka*)	Douglas' Pouch
130217	Douglas か (Douglas-*ka*)	Douglas' Pouch
130217	ダグラス窩 (*Dagurasu-ka*)	pouch of Douglas
130217	ダグラスか (*Dagurasu-ka*)	pouch of Douglas
130217	直腸子宮窩 (*chokuchō-shikyū-ka*)	rectouterine pouch
130217	直腸子宮か (*chokuchō-shikyū-ka*)	rectouterine pouch

Table 4: A part of the lexicon used

Each bilingual synonym set consists of items that have the same synset ID. 部分 (*bubun*) is a synonym of 部位 (*bui*). か (*ka*) is a *hiragana* representation of 窩 (*ka*). ダグラス (*Dagurasu*) is a Japanese transcription of 'Douglas'.

recall R and F-score F defined as follows:

$$P = \frac{C}{T}, R = \frac{C}{N}, F = \frac{2PR}{P+R}, \quad (5)$$

where C, T and N are the number of correctly predicted pairs as synonyms, predicted pairs to become synonyms, and synonym pairs in the lexicon[9], respectively.

We compared the results with the baseline and the upper bound. The baseline assumes that each bilingual lexical item is a bilingual synonym if either the Japanese or English terms are identical. The upper bound assumes that all the monolingual synonyms are known and each bilingual item is a bilingual synonym if the Japanese terms and the English terms are synonymous. The baseline represents the performance when we do not consider spelling variations, and the upper bound shows the limitation of the monolingual approach.

4.3 Result

Table 6 shows the evaluation scores of our experiments. The 'monolingual' and 'bilingual' methods are described in Section 3.2.1. We obtained high precision and recall scores, although we used features primarily with spelling variations. Both methods significantly outperform the baseline, and show the importance of considering spelling variations.

[9] N includes the number of synonym pairs filtered out from training set by the bigram similarity threshold T_b.

Set	Method	Precision	Recall	F-score
dev.	baseline	0.977 (31845/32581)	0.410 (31845/77706)	0.577
	monolingual	**0.911** (74263/81501)	**0.956** (74263/77706)	**0.932**
	bilingual	0.879 (72782/82796)	0.937 (72782/77706)	0.907
	upper bound	0.984 (77706/78948)	1	0.992
test	baseline	0.972 (33382/34347)	0.392 (33382/85091)	0.559
	monolingual	**0.900** (79099/87901)	**0.930** (79099/85091)	**0.914**
	bilingual	0.875 (77640/88714)	0.912 (77640/85091)	0.893
	upper bound	0.979 (85091/86937)	1	0.989

Table 6: Evaluation scores

The 'monolingual' method achieved higher precision and recall than the 'bilingual' method. It indicates that monolingual synonym identification is effective in finding bilingual synonyms. The upper bound shows that there are still a few errors by the assumption used by the 'monolingual' method. However, the high precision of the upper bound represents the well-formedness of the lexicon we used. We need more experiments on other bilingual lexicons to conclude that our method is available for

Features	Precision	Recall	F-score
All	0.911	0.956	0.932
$-h_{1F}, h_{1E}$	0.911	**0.974**	**0.941**
$-h_{2F}, h_{2E}$	0.906	0.947	0.926
$-h_{3F}, h_{3E}$	0.939	0.930	0.934
$-h_{4F}, h_{4E}$	0.919	0.734	0.816
$-h_{5F}, h_{5E}$	0.869	0.804	0.831
$-h_6, h_7$	**0.940**	0.934	0.937
$-$combs.	0.936	0.929	0.932

Table 7: Evaluation scores of the bilingual method with removing features on the development set
$-h$ represents removing the feature h and combinatorial features using h. $-$combs. represents removing all the combinatorial features.

many kinds of lexicons.

To investigate the effectiveness of each feature, we compared the scores when we remove several features. Table 7 shows these results. Contrary to our intuition, we found that features of agreement of the first characters (h_1) remarkably degraded the recall without gains in precision. One of the reasons for such results is that there are many cases of non-synonyms that have the same first character. We need to investigate more effective combinations of features or to apply other machine learning techniques for improving the performance. From these results, we consider that the features of h_4 are effective for improving the recall, and that the features of h_2 and h_5 contribute improvement of both the precision and the recall. h_3, h_6, h_7, and combinatorial features seem to improve the recall at the expense of precision. Which measure is important depends on the importance of our target for using this technique. It depends on the requirements that we emphasize, but in general the recall is more important for finding more bilingual synonyms.

5 Conclusion and future work

This paper proposed a method for identifying bilingual synonyms in a bilingual lexicon by using clues of spelling variations. We described the notion of bilingual synonyms, and presented two approaches for identifying them: one is to directly predict the relation, and another is to merge monolingual synonyms identified, according to the bilingual lexicon. Our experiments showed that the proposed method significantly outperformed the method that did not use features primarily with spelling variations; the proposed method extracted bilingual synonyms with high precision and recall. In addition, we found that merging monolingual synonyms by the dictionary is effective for finding bilingual synonyms; there occur few errors through the assumption described in Section 3.2.1.

Our future work contains implementing more features for identifying synonymous relations, constructing bilingual synonym sets, and evaluating our method for specific tasks such as thesaurus construction or cross-lingual information retrieval.

Currently, the features used do not include other clues with spelling variations, such as the weighted edit distance, transformation patterns, stemming and so on. Another important clue is distributional information, such as the context. We can use both monolingual and bilingual corpora for extracting distributions of terms, and bilingual corpora are expected to be especially effective for our goal.

We did not perform an experiment to construct bilingual synonym sets from synonym pairs in this paper. Described in Section 3.1, bilingual synonym sets can be constructed from bilingual synonym pairs by assuming some approximations. The approximation that permits transitivity of bilingual synonymous relations increases identified bilingual synonyms, and thus causes an increase in recall and decrease in precision. It is an open problem to find appropriate strategies for constructing bilingual synonym sets.

Finally, we plan to evaluate our method for specific tasks. For data-driven machine translation, it is expected that data sparseness problem is alleviated by merging the occurrences of low-frequency terms. Another application is cross-lingual information retrieval, which can be improved by using candidate expanded queries from bilingual synonym sets.

Acknowledgments

This work was partially supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan) and Japanese/Chinese Machine Translation Project in Special Coordination Funds for Promoting Science and Technology (MEXT, Japan). We thank

Japan Science and Technology Agency (JST) for providing a useful bilingual lexicon with synonymous information. We acknowledge the anonymous reviewers for helpful comments and suggestions.

References

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proc. of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Vincent D. Blondel, Anahí Gajardo, Maureen Heymans, Pierre Senellart, and Paul Van Dooren. 2004. A measure of similarity between graph vertices: Applications to synonym extraction and web searching. *SIAM Review*, 46(4):647–666.

Carolyn J. Croach and Bokyung Yang. 1992. Experiments in automatic statistical thesaurus construction. In *Proc. of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 77–88. ACM Press.

James R. Curran. 2004. *From Distributional to Semantic Similarity*. Ph.D. thesis, University of Edinburgh.

Béatrice Daille, Benoît Habert, Christian Jacquemin, and Jean Royauté. 1996. Empirical observation of term variations and principles for their description. *Terminology*, 3(2):197–258.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers.

Donald Hindle. 1990. Noun classification from predicate-argument structures. In *Proc. of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese morphological analyser JUMAN. In *Proc. of International Workshop on Sharable Natural Language Resources*, pages 22–28.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. 2003. Identifying synonyms among distributionally similar words. In *Proc. of the 2003 International Joint Conference on Artificial Intelligence*, pages 1492–1493.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. of the 17th International Conference on Computational Linguistics*, volume 2, pages 768–774.

Julie B. Lovins. 1968. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31.

Takeshi Masuyama, Satoshi Sekine, and Hiroshi Nakagawa. 2004. Automatic construction of Japanese KATAKANA variant list from large corpus. In *Proc. of the 20th International Conference on Computational Linguistics*, volume 2, pages 1214–1219.

Philippe Muller, Nabil Hathout, and Bruno Gaume. 2006. Synonym extraction using a semantic distance on a dictionary. In *Proc. of TextGraphs: the 2nd Workshop on Graph Based Methods for Natural Language Processing*, pages 65–72.

Young C. Park and Key-Sun Choi. 1996. Automatic thesaurus construction using Bayesian networks. *Information Processing and Management*, 32(5):543–553.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Yonggang Qiu and Hans-Peter Frei. 1993. Concept-based query expansion. In *Proc. of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, pages 160–169.

Ariel S. Schwartz and Marti A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proc. of the 8th Pacific Symposium on Biocomputing*, pages 451–462.

Mitsuo Shimohata and Eiichiro Sumita. 2002. Automatic paraphrasing based on parallel corpus for normalization. In *Proc. of the 3rd International Conference on Language Resources and Evaluation*, volume 2, pages 453–457.

Scott A. Waterman. 1996. Distinguished usage. In *Corpus Processing for Lexical Acquisition*, pages 143–172. MIT Press.

Hua Wu and Ming Zhou. 2003. Optimizing synonym extraction using monolingual and bilingual resources. In *Proc. of the 2nd International Workshop on Paraphrasing*.