Improving Calculation of Contextual Similarity for Constructing a Bilingual Dictionary via a Third Language

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

Abstract

A novel method is proposed for measuring contextual similarity by "weighted overlapping ratio (WOR)" to construct a bilingual dictionary of a new language pair from two bilingual dictionaries sharing one language. The WOR alleviates the effect of a noisy seed dictionary resulting from merger of two bilingual dictionaries via a third language. Combined use of two word-association measures for extracting contexts from corpora is also proposed to compensate their weaknesses. Experiments on constructing a Japanese-Chinese dictionary via English show that the proposed method outperforms the conventional method based on cosine similarity.

1 Introduction

With the growth of communication via the Internet, people have more chance to access documents written in various languages. The number of Internet users in the Arabic, Russian, and Chinese languages have increased at least tenfold times in the last decade. It was sufficient in the past to bridge the gap between English and another language by using bilingual resources and services. For directly accessing Web contents written in various languages, however, multilingual dictionaries, translation, and information retrieval are required.

The present study focuses on a so-called triangulation approach for constructing a bilingual dictionary by merging two bilingual dictionaries sharing one language. For example, if Chinese-to-English and English-to-Arabic dictionaries are available, a Chinese-to-Arabic dictionary can be derived by collecting pairs of Chinese and Arabic terms (hereafter, "term pairs") that have common English translations. A serious problem with this approach, however, is how to filter out false term pairs, caused by polysemy of English terms. To validate each term pair as translations, we calculate the context similarity between the terms on the basis of the distributional hypothesis (Harris, 1954).

In triangulation approach, calculation of the context similarity in different languages is required. Previous studies have calculated context similarities by context vector projection onto another language by using a seed bilingual dictionary. Though this approach is effective, translation perplexity caused by context vector projection may cause negative effects described in Section 3.3. Instead of projection, our proposed method avoids this problem by using a "weighted overlapping ratio," which directly maps words in context vectors in different languages.

2 Related work

Tanaka and Umemura (1994) proposed a triangulation method of constructing a bilingual dictionary. Their method has been augmented by using semantic classes (Bond et al., 2001) and parts of speech and cognates (Zhang et al., 2005).

Several methods of constructing a bilingual dictionary from contextual similarity have been proposed (Rapp, 1995; Kaji and Aizono, 1996; Tanaka and Iwasaki, 1996; Fung and Yee, 1998; Rapp, 1999; Sammer and Soderland, 2007). Most of them are based on context vector projection. Rapp (1999) replaced a word in the context vector with the translation first appeared in the dictionary, while Fung and Yee (1998) gave each translation a weight inversely proportional to the order of the translation in the dictionary. As another provision for translating context vectors,

International Joint Conference on Natural Language Processing, pages 1057–1061,
Nagoya, Japan, 14-18 October 2013.

mutual projection of context vectors was proposed (Fišer et al., 2011). Adapting a seed bilingual dictionary to the domain of comparable corpora has been proved to be effective (Kaji, 2005; Morin and Prochasson, 2011).

Other approaches (Déjean and Gaussier, 2002; Daille and Morin, 2005; Hazem et al., 2011) proposed methods based on identification of second-order affinities. Kaji et al. (2008) created a correlation matrix of context words versus translations. Vulić and Moens (2012) proposed a bilingual LDA model in which the term pairs are obtained on the basis of similar distributions of language-independent latent topics.

3 Proposed method

The proposed method is overviewed in Figure 1. Each step of the proposed method is described in the following subsections.

3.1 Merging bilingual dictionaries

It is supposed that a bilingual dictionary between a source language S and a target language T can be constructed via a third language P. A bilingual dictionary, $D_{L,L'}$, between two languages, L and L', can be defined as a set of term pairs $\{(w_l, w_{l'})\} \subseteq L \times L'$, where term $w_l \in L$ can be translated as term $w_{l'} \in L'$.[1]

It is assumed that two bilingual dictionaries, $D_{S,P}$ and $D_{P,T}$, are available. The merged bilingual dictionary between S and T, namely, $D_{S,T}$, is obtained from

$$\{(w_s, w_t) | \exists w_p: (w_s, w_p) \in D_{S,P} \wedge (w_p, w_t) \in D_{P,T}\}.$$

Note that term w_s cannot necessarily be translated into term w_t because of polysemy of term w_p. Such term pair (w_s, w_t) makes “noise” in the merged dictionary.

$D_{S,T}$ is used as a candidate set of term pairs to be ranked. It is also used as a seed bilingual dictionary to calculate the similarity of contexts in languages S and T.

3.2 Extracting context

Spurious term pairs in the merged bilingual dictionary should be removed. The similarity of senses of each term pair is estimated by comparing their contexts. We represent the context of term w as a *weighted set of associated words*, i.e., words that are semantically or topically related with w. The weighted set of associated words, $C(w)$, is denoted as $\{w_1/\alpha_1,\ w_2/\alpha_2,\ ...,\ w_K/\alpha_K\}$ where w_k is an associated word of w, and α_k is its weight assigned by the word-association measure based on their co-occurrence frequencies.

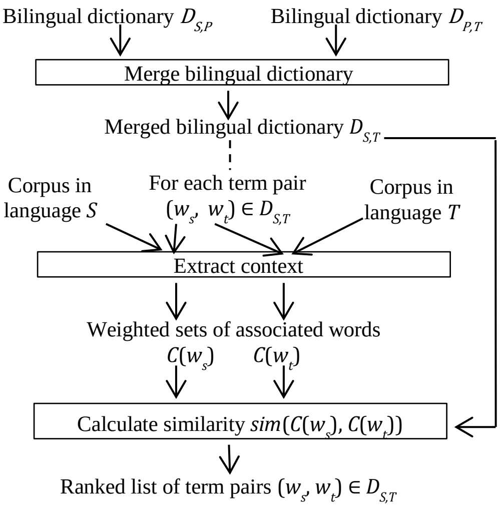

Bilingual dictionary $D_{S,P}$
Bilingual dictionary $D_{P,T}$
Merge bilingual dictionary
Merged bilingual dictionary $D_{S,T}$
Corpus in language S
For each term pair $(w_s, w_t) \in D_{S,T}$
Corpus in language T
Extract context
Weighted sets of associated words
$C(w_s)$
$C(w_t)$
Calculate similarity $sim(C(w_s), C(w_t))$
Ranked list of term pairs $(w_s, w_t) \in D_{S,T}$

Figure 1. Overview of proposed method

(1) Using a single word-association measure

Among words that co-occur with w in the corpus, only words that have association measure scores in the top-M%[2] are kept as associated words. The word-association measures employed are log-likelihood ratio (LLR), pointwise mutual information (MI), chi-squared score (χ^2), and discounted log-odds ratio (LOR) (Laroche and Langlais, 2010).

(2) Using combination of word-association measures

Each association measure has its own weakness in capturing word association. For example, LLR tends to overestimate frequent words, while MI tends to do infrequent ones. In general, infrequent associated words have less possibility to be matched when comparing two sets of associated words. A combination of these measures is expected to compensate for each weakness.

Associated words whose first association measure is in the top-M_1% and second is in the top-M_2% are kept. A weight corresponding to each associated word is given by the second measure. Two kinds of combinations are considered: LLR-MI and LLR-LOR, which respectively represent the first and second measures.

3.3 Calculating similarity between weighted sets of associated words

[1] We describe that the vocabulary set of a language L also as L in short.

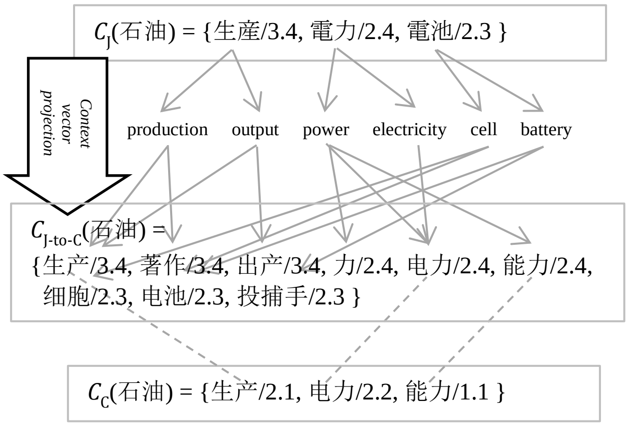
[2] A fixed threshold value of the association score was not used for keeping associated words because the proposed method obtained better results in our experiment.

A problematic case of context vector projection is illustrated in Figure 2. For calculating contextual similarity, such as a cosine, the context vectors[3] must be projected onto associated-word dimensions in the same language. In this approach, associated words are duplicated by translation perplexity. In this example, each word associated with the Japanese word "石油" sekiyu 'petroleum' has several possible English translations. It yields unnecessary Chinese associated words such as "力" li 'power' and "细胞" xibao 'cell (in the biological sense),' and then falsely decreases the cosine value because the norm of the projected vector increases.[4]

To avoid this problem, two sets of associated words are directly compared. For two weighted sets of associated words, $C(w_s) = \{w_k/\alpha_k\}$ and $C(w_t) = \{w'_l/\alpha'_l\}$, a *weighted overlapping ratio* (WOR) is defined as:

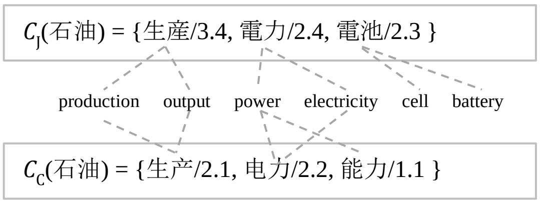
$$sim(C(w_s), C(w_t)) = \frac{1}{2}\left\{\frac{\sum_{k\in P}\alpha_k}{\sum_k \alpha_k} + \frac{\sum_{l\in Q}\alpha'_l}{\sum_l \alpha'_l}\right\}$$

where $P = \{k|\exists w'_l\text{: } (w_k, w'_l) \in D_{S,T}\}$, $Q = \{l|\exists w_k\text{: } (w'_l, w_k) \in D_{T,S}\}$, and $D_{S,T}$ and $D_{T,S}$ are seed bilingual dictionaries between languages S and T. Term pairs (w_s, w_t) in the merged dictionary are ranked in order of their WORs. An example calculation of WOR is shown in Figure 3. The side effect from a noisy seed dictionary is considered to be moderated, since an unnecessary term is added only once per noisy term.


C_J(石油) = {生産/3.4, 電力/2.4, 電池/2.3 }
Context vector projection
production
output
power
electricity
cell
battery
$C_{J\text{-to-}C}$(石油) =
{生产/3.4, 著作/3.4, 出产/3.4, 力/2.4, 电力/2.4, 能力/2.4, 细胞/2.3, 电池/2.3, 投捕手/2.3 }
C_C(石油) = {生产/2.1, 电力/2.2, 能力/1.1 }

$$sim_{\cos}\left(C_J(\text{石油}), C_C(\text{石油})\right) = \cos\left(C_{J-\text{to}-C}(\text{石油}), C_C(\text{石油})\right) = 0.565$$

Figure 2. Calculation of contextual similarity by context-vector projection


C_J(石油) = {生産/3.4, 電力/2.4, 電池/2.3 }
production
output
power
electricity
cell
battery
C_C(石油) = {生产/2.1, 电力/2.2, 能力/1.1 }

$$sim_{\mathrm{WOR}}(C_J(\text{石油}), C_C(\text{石油})) = \frac{1}{2}\left(\frac{3.4+2.4}{3.4+2.4+2.3} + \frac{2.1+2.2+1.1}{2.1+2.2+1.1}\right) = 0.858$$

Figure 3. Calculation of WOR

4 Experiments

Experiments on constructing a Japanese-Chinese bilingual dictionary via English as a third language were carried out. Note that this approach is applicable for any language combinations. We report three kinds of comparison: WOR vs. cosine similarity, word-association measures, and newspaper corpus vs. Wikipedia corpus.

Window size W for counting co-occurrence frequencies was fixed to 10. Five-fold cross validation was conducted for optimizing parameters for choosing associated words (M; M_1 and M_2).[5] The final evaluation score is output by averaging all five results.

4.1 Experimental settings

Two sets of comparable corpora were employed for our experiment: newspaper and Wikipedia. The newspaper set consists of the Mainichi Shimbun Corpus (2000-2010; 22.3GB) and the Xinhua News Corpus in LDC Chinese Gigaword (2000-2010; 4.24GB). The Wikipedia set consists of Japanese and Chinese Wikipedia articles dumped on September 2012 (Japanese: 821k articles; 3.1GB / Chinese: 520k articles; 0.7GB).

EDR Japanese-to-English/English-to-Japanese dictionaries and LDC Chinese-to-English dictionary[6] were used as input dictionaries. Each EDR dictionary has 376k word pairs, including 161k English distinct words and 221k Japanese distinct words. The LDC dictionary consists of 82k distinct word pairs.

3,000 term pairs (each term occurs at least 2,500 times in the corpus) were randomly extracted as the test data from the merged Japanese-Chinese dictionary. Each term pair was labeled as *translation* or *non-translation* by ma-

[3] The weighted set of associated words is compatible to the context vector with dimensions of associated words in the vector space model.

[4] Rapp (1999) substituted each associated word to only a single translation. In that case, however, a noisy seed dictionary significantly decreases the probability that the translation is appropriate.

[5] The optimized values were as follows: M = 1.02 (%) for LOR and M_1 = 13.5, M_2 = 9.4 (%) for MI-LLR.

[6] The English-to-Chinese dictionary was obtained by inversing the LDC dictionary.

jority decision of three human annotators. The test set consists of 1,053 *translation* pairs and 1,947 *non-translation* pairs.

Precision P and recall R for the term pairs with the top-δ% of WORs or cosine values were calculated. Some *translation* term pairs could not be correctly judged because those terms are sometimes only used for representing different senses from a sense in the comparable corpus. The recall therefore does not reach a higher value compared with the precision value. For this reason, an F_β-score with $\beta = 0.5$ was adopted as the final evaluation score so as to emphasize precision as twice as important as recall. The best $F_{0.5}$-scores were obtained when $\delta = 20$ (%) (see Table 1).

4.2 WOR vs. cosine similarity

To confirm the effect of WOR, it was compared with the conventional cosine by context vector projection. The best evaluation scores are listed in Table 2. WOR outperformed the cosine measure on both corpus sets.

The merged dictionary for comparing associated words can also be substituted by existing bilingual dictionaries between languages S and T if available. To examine the effect of using the noisy seed bilingual dictionary, additional experiments in using the EDR Japanese-Chinese dictionary (223k term pairs) as a seed dictionary were conducted. The $F_{0.5}$-score by WOR with this setting was 0.743, while 0.721 by cosine measure. The drop in the $F_{0.5}$-score by using the merged dictionary as the seed were 1.4 points by WOR, which were smaller than the drop (3.0 points) obtained by the cosine. This result shows that WOR is more robust than the cosine in the case that a noisy seed dictionary is used.

4.3 Single measure vs. combination of measures

Experiments on using all word association measures were carried out. Among the single word-association measure, the highest $F_{0.5}$-score of 0.689 was obtained by LOR as listed in Table 2, and it confirmed a previous comparative experiment (Laroche and Langlais, 2010). Both combinations of the multiple association measures outperformed LOR on the newspaper set. These results indicate that the weakness that LOR covered could also be covered by LLR.

4.4 Newspaper vs. Wikipedia as the comparable corpus

The characteristics of the results obtained from each comparable corpus set described in Section 4.1 were examined. As listed in Table 2, combinations of the multiple measures did not achieve significant improvement on the Wikipedia corpus set. Meanwhile, the overall performance of Wikipedia did not significantly degrade in comparison with the newspaper set, although larger corpora give more appropriate associated word sets for each term pair. One reason for that result is the high comparability of Wikipedia data.

δ (%)	P	$F_{0.5}$
10	0.908	0.611
20	0.833	**0.729**
30	0.744	0.723
40	0.640	0.659
50	0.567	0.605

Table 1. Evaluation scores attained some values of δ (%) (settings: WOR, LLR-MI, newspaper)

	Corpus set	Newspaper		Wikipedia	
	Measure	P	$F_{0.5}$	P	$F_{0.5}$
WOR	LLR	0.721	0.631	0.664	0.646
	MI	0.704	0.616	0.711	0.691
	LOR	0.788	0.689	0.792	0.693
	χ^2	0.717	0.622	0.717	0.632
	LLR-MI	**0.833**	**0.729**	**0.796**	**0.696**
	LLR-LOR	0.829	0.725	0.708	0.688
cosine	LLR	0.622	0.609	0.708	0.531
	LLR-MI	0.783	0.691	0.775	0.684

Table 2. Evaluation scores

5 Concluding remarks

A novel method for constructing bilingual dictionaries via a third language is proposed. It applies a novel context similarity criterion, namely, a “weighted overlapping ratio” (WOR) for alleviating negative effects from translation perplexity. In addition, a method for combining word-association measures is developed. Experiments demonstrated the effectiveness of the proposed method: the proposed method achieved the highest $F_{0.5}$-score 0.729, thereby outperforming the $F_{0.5}$-score 0.691 by the conventional cosine-similarity method in the case of projecting context vectors onto English.

A future direction is applying word-sense-disambiguation techniques to associated words. By separating polysemous associated words into some classes corresponding to each sense, we could avoid the negative effect from unrelated senses of the associated words.

Acknowledgments

This work was partially supported by Grant-in-Aid for Scientific Research, MEXT (22300032). We thank Prof. Masanori Kato for proofreading.

References

Bond, Francis, Ruhaida Binti Sulong, Takefumi Yamazaki, and Kentaro Ogura. 2001. Design and construction of a machine-tractable Japanese-Malay dictionary. In *Proceedings of MT Summit XIII*, pages 53–58.

Daille, Béatrice and Emmanuel Morin. 2005. French-English terminology extraction from comparable corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, Vol. 3651, pages 707–718.

Déjean, Hervé and Éric Gaussier. 2002. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, :1–22.

Fišer, Darja, Špela Vintar, Nikola Ljubešić, and Senja Pollak. 2011. Building and using comparable corpora for domain-specific bilingual lexicon extraction. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 19–26.

Fung, Pascale and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting on Association for Computational Linguistics*, Vol. 1, pages 414–420.

Harris, Zellig. 1954. Distributional structure. *Word*, 10(23):146–162.

Hazem, Amir, Emmanuel Morin, and Sebastian Peña Saldarriaga. 2011. Bilingual lexicon extraction from comparable corpora as metasearch. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, pages 35–43.

Kaji, Hiroyuki. 2005. Extracting translation equivalents from bilingual comparable corpora. *IEICE Transactions on Information and Systems*, E88-D(2):313–323.

Kaji, Hiroyuki and Toshiko Aizono. 1996. Extracting word correspondences from bilingual corpora based on word co-occurrence information. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 23–28.

Kaji, Hiroyuki, Shin'ichi Tamamura, and Dashtseren Erdenebat. 2008. Automatic construction of a Japanese-Chinese dictionary via English. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 699–706.

Laroche, Audrey and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 617–625.

Morin, Emmanuel and Emmanuel Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 27–34.

Rapp, Reinhard. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, pages 320–322.

Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 519–526.

Sammer, Marcus and Stephen Soderland. 2007. Building a sense-distinguished multilingual lexicon from monolingual corpora and bilingual lexicons. In *Proceedings of MT Summit XI*, pages 399–406.

Tanaka, Kumiko and Hideya Iwasaki. 1996. Extraction of lexical translations from non-aligned corpora. In *Proceedings of the 16th Conference on Computational Linguistics*, Vol. 2, pages 580–585.

Tanaka, Kumiko and Kyoji Umemura. 1994. Construction of a Bilingual Dictionary Intermediated by a Third Language. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 297–303.

Vulić, Ivan and Marie-Francine Moens. 2012. Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 449–459.

Zhang, Yujie, Qing Ma, and Hitoshi Isahara. 2005. Construction of a Japanese-Chinese bilingual dictionary using English as an intermediary. *International Journal of Computer Processing of Oriental Languages*, 18(1):23–39.