Automatic Construction of a Japanese-Chinese Dictionary via English

Hiroyuki Kaji, Shin'ichi Tamamura, and Dashtseren Erdenebat
Department of Computer Science, Shizuoka University
3-5-1 Johoku, Naka-ku, Hamamatsu-shi, 432-8011, Japan
E-mail: kaji@inf.shizuoka.ac.jp

Abstract

This paper proposes a method of constructing a dictionary for a pair of languages from bilingual dictionaries between each of the languages and a third language. Such a method would be useful for language pairs for which wide-coverage bilingual dictionaries are not available, but it suffers from spurious translations caused by the ambiguity of intermediary third-language words. To eliminate spurious translations, the proposed method uses the monolingual corpora of the first and second languages, whose availability is not as limited as that of parallel corpora. Extracting word associations from the corpora of both languages, the method correlates the associated words of an entry word with its translation candidates. It then selects translation candidates that have the highest correlations with a certain percentage or more of the associated words. The method has the following features. It first produces a domain-adapted bilingual dictionary. Second, the resulting bilingual dictionary, which not only provides translations but also associated words supporting each translation, enables contextually based selection of translations. Preliminary experiments using the EDR Japanese-English and LDC Chinese-English dictionaries together with Mainichi Newspaper and Xinhua News Agency corpora demonstrate that the proposed method is viable. The recall and precision could be improved by optimizing the parameters.

1. Introduction

The need for machine translation and cross-language information retrieval is growing for a variety of language pairs, including those for which wide-coverage bilingual dictionaries are not available. Therefore, we need to develop a method of constructing a bilingual dictionary for new language pairs from available language resources. Since bilingual dictionaries between such pairs of languages and a third language, usually English, are often available, it is normal to combine them into a bilingual dictionary for the new language pair. This method, however, suffers from spurious translations caused by the ambiguity of intermediary third-language words.

To solve this problem, Tanaka and Umemura (1994) proposed a method of distinguishing between correct and incorrect translations based on the assumption that the more intermediary third-language words there were the more likely a correct translation would result. Their method has been augmented by using additional clues such as semantic classes of words (Bond, et al., 2001) and parts of speech and constituent characters of words (Zhang, et al., 2007). A variant of Tanaka and Umemura's method was also proposed by Shirai and Yamamoto (2001). However, how to eliminate spurious translations still remains an unsolved problem. We propose a novel method that uses a pair of monolingual corpora from two languages in the same domain to eliminate spurious translations from a combination of two bilingual dictionaries. Note that such pairs of monolingual corpora are available for many language pairs and in many domains.

2. Proposed method

Our proposed method is based on an iterative algorithm for correlating the associated words of a first-language word with its second-language translations, which we originally developed as a means of unsupervised word sense disambiguation (Kaji and Morimoto, 2002). The algorithm calculates a correlation matrix of associated words versus translations for each first-language word from a bilingual dictionary and two monolingual corpora, the first in the first language and the other in the second language. When it is used with a bilingual dictionary containing spurious translations, the algorithm allows spurious translations to be eliminated from the bilingual dictionary because they are likely to have high correlations with very few of the associated words. Note that translations not used in the domain of the corpora also have high correlations with few of the associated words. Precisely speaking, the proposed method produces a "domain-adapted" bilingual dictionary.

The proposed method consists of the following steps as outlined in Fig. 1, where we have assumed that the first, second, and third languages are Japanese, Chinese, and English.

(1) Combine Japanese-English and Chinese-English dictionaries into a Japanese-Chinese dictionary.
(2) Extract word associations from both Japanese and Chinese corpora.
(3) Align Japanese word associations with Chinese word associations and, for each Japanese entry word, calculate a correlation matrix of Japanese associated words versus Chinese translations.
(4) Select Chinese translations supported by a certain percentage or more of the Japanese associated words.

2.1 Combining bilingual dictionaries

A Chinese word is regarded as a translation of a Japanese word when they have one or more English translations in common. Note that the resultant Japanese-Chinese dictionary is "noisy." An example is given in Fig. 2; the Japanese entry word "工場" (*factory*, *plant*) not only has correct Chinese translations such as "厂" and "工场" but also spurious translations such as "植株" (*flora*) and "作

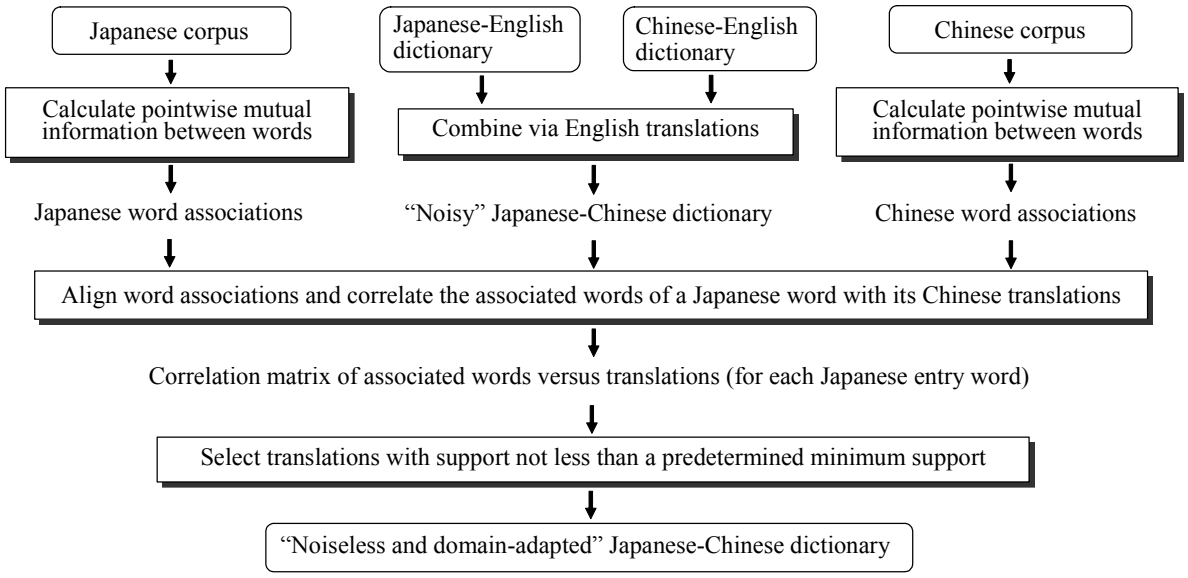


Fig. 1: Overview of proposed method

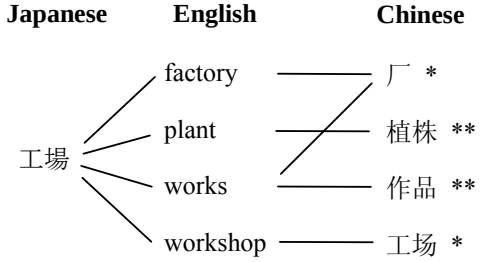
品" (*piece of work*) caused by the ambiguity of the intermediary English words "plant" and "works."

2.2 Extracting word associations

Word associations, i.e., pairs of words having mutual information not less than a predetermined threshold, are extracted from both Japanese and Chinese corpora. The mutual information between words x and x' is defined as:

$$MI(x, x') = \log \frac{\Pr(x, x')}{\Pr(x) \cdot \Pr(x')}, \quad [1]$$

where Pr(x) denotes the occurrence probability of x, and Pr(x, x') the co-occurrence probability of x and x' (Church and Hanks, 1990). Occurrence probabilities are estimated by counting the frequency of each word occurring in a corpus and co-occurrence probabilities are estimated by counting the frequency of each pair of words co-occurring in a window. We adopted a medium-sized window that covered a few sentences since we were interested in pairs of topically associated words. For example, Japanese word associations such as (工場, バルブ (*valve*)), (工場, 工程 (*manufacturing process*)), (工場, 飼料 (*feed, fodder*)), and (工場, 製造 (*manufacture*)) would be extracted from a Japanese corpus.



*Correct translation, **Incorrect translation

Fig. 2: Combining Japanese word and Chinese words via English translations

2.3 Correlating associated words with translations

Japanese word associations are aligned with Chinese word associations by consulting a "noisy" Japanese-Chinese dictionary. Note that not all Japanese word associations can be aligned successfully with their Chinese counterparts because of disparity in topical coverage between the Japanese and Chinese corpora as well as incomplete coverage of the Japanese-Chinese dictionary. Note also that a Japanese word association can often be aligned with two or more Chinese word associations, and vice versa.

To cope with both alignment failure and ambiguity in alignment, a correlation matrix of associated words versus translations is calculated iteratively for each Japanese entry word. The correlation between the i-th associated word $x'(i)$ of entry word x and its j-th translation $y(j)$ is defined as:

$$C_n(x'(i), y(j)) = MI(x'(i), x) \cdot \frac{C'_n(x'(i), y(j)) + \alpha \cdot C''_n(x'(i), y(j))}{\max_k [C'_n(x'(i), y(k)) + \alpha \cdot C''_n(x'(i), y(k))]}, \quad [2]$$

where the suffix, n, denotes iteration number, and α is a parameter specifying the relative weight between primary and secondary correlation factors C' and C''.

$$C'_n(x'(i), y(j)) = \sum_{x'' \in A(x, x'(i))} C_{n-1}(x'', y(j)), \quad [3]$$

where $A(x, x'(i))$ denotes a set consisting of words each of which is associated with both x and $x'(i)$.

$$C''_n\left(x'(i), y(j)\right) = \max_{y'}\left(MI(y', y(j)) \cdot \sum_{x'' \in B((x,x'(i)),(y(j),y'))} C_{n-1}\left(x'', y(j)\right)\right), \quad [4]$$

where $B((x, x'(i)), (y(j), y'))$ denotes a set consisting of words each of which is associated with both x and $x'(i)$ and has at least one translation associated with both $y(j)$ and y' where word association $(x, x'(i))$ can be aligned with word association $(y(j), y')$.

The primary and secondary correlation factors, both of which are defined by using the correlations between other associated words and the same translation, respectively overcome alignment failure and ambiguity in alignment. See the Appendix for the underlying assumptions from which the above formulas were derived. A correlation matrix of associated words versus translations is calculated iteratively with the following initial values:

$$C_0\left(x'(i), y(j)\right) = MI\left(x'(i), x\right). \quad [5]$$

Also see the Appendix for how the correlation values converge. Figure 3(a) shows part of a correlation matrix calculated for the Japanese entry word "工場."

	厂	植株	作品	工场
バルブ	3.98	0.81	0.23	3.75
工程	3.18	0.45	0.58	3.45
飼料	0.54	3.32	0.17	1.18
製造	2.89	0.43	0.61	3.13
:	:	:	:	:

(a) Correlation matrix for "工場"

	厂	植株	作品	工场
バルブ	1	0	0	0
工程	0	0	0	1
飼料	0	1	0	0
製造	0	0	0	1
:	:	:	:	:
Support	0.03	0.04	0.00	0.93

(b) Binary matrix and support of translations

Fig. 3: Correlation matrix and support of translations

2.4 Selecting translations

We define the "support" of a translation as the proportion of associated words having the highest correlations with it. To calculate the support of each translation, the correlation matrix of associated words versus translations is converted into a binary matrix; the cell whose value is largest within a row is set to 1, and the other cells are set to 0. Then, the support of a translation is calculated as the number of 1's in the corresponding column divided by the number of rows. Finally, translations with a support of not less than a predetermined "minimum support" are selected. Thus, the translations contained in the noisy Japanese-Chinese dictionary are screened. A binary matrix is shown in Fig. 3(b) together with supports for translations. The spurious translations "植株" and "作品" are eliminated when the minimum support is set to 0.1. Note that the correct translation "厂" is also eliminated because it is probably less dominant in the Chinese corpus than another correct translation "工场."

3. Preliminary experiments

3.1 Experimental setting

We did experiments that focused on nouns by using the EDR (Japan Electronic Dictionary Research Institute) Japanese-English Bilingual Dictionary and the LDC (Linguistic Data Consortium) Chinese-English Translation Lexicon. Both Japanese-Chinese and Chinese-Japanese noun dictionaries were produced, and they were screened by using a Japanese corpus consisting of Mainichi newspaper articles (2000/01-2005/12, 632 Mbytes) and a Chinese corpus consisting of Xinhua News Agency's newswire articles (1999/01-2004/12, 473 Mbytes).

Noun associations were extracted from both the Japanese and Chinese corpora as follows. First, the Japanese texts were segmented into words by using JUMAN and the Chinese texts were segmented into words by using a Chinese morphological analyzer developed by Nakagawa and Uchimoto (2007). Then, Japanese nouns occurring not less than 100 times and Chinese nouns occurring not less than 30 times were extracted. Furthermore, pairs of nouns co-occurring in a window that covered the preceding 25 and succeeding 25 content words were extracted, and their respective frequencies were counted. Finally, pairs of nouns having mutual information of not less than 1.0 were extracted.

The extracted noun associations were used together with the noisy Japanese-Chinese noun dictionary to calculate a correlation matrix of Japanese associated nouns versus Chinese translations for each of the extracted Japanese nouns. They were also used together with the noisy Chinese-Japanese noun dictionary to calculate a correlation matrix of Chinese associated nouns versus Japanese translations for each of the extracted Chinese nouns. It should be noted that, for computational reasons, the number of associated nouns in a matrix was restricted to 700 or less in descending order of mutual information value.

A total of 8,284 Japanese nouns out of 10,003 occurring not less than 100 times had two or more translation candidates in the noisy Japanese-Chinese dictionary. The

translation candidates of these Japanese entry nouns were screened by using the correlation matrices of Japanese associated nouns versus Chinese translations. Likewise, a total of 8,426 Chinese nouns out of 9,288 occurring not less than 30 times had two or more translation candidates in the noisy Chinese-Japanese dictionary. The translation candidates of these Chinese entry nouns were screened by using the correlation matrices of Chinese associated nouns versus Japanese translations.

3.2 Experimental results

It is obvious that the screening results varied with minimum support. When it was set to 0.1, the number of Chinese translations per Japanese entry noun, averaged over the 8,284 entry nouns, was reduced from 15.77 to 1.40. Likewise, the number of Japanese translations per Chinese entry noun, averaged over the 8,426 entry nouns, was reduced from 27.09 to 1.33. This drastic reduction is due to the nature of the proposed method in that it eliminates not only spurious translations but also translations not used in the domain of the corpora.

Table 1(a) lists examples from the results of screening the Japanese-Chinese dictionary; for each entry noun, translation candidates are listed in descending order of support, and translations with support not less than the minimum support (0.1) are boldfaced.

- For entry noun "ホール," which is a transliteration of both English words "*hall*" and "*hole*," a correct translations "大厅" (*hall*) was selected but another correct translation "洞" (*hole*) was eliminated; although "洞" was supported by "オゾン" (*ozone*), "量子力学" (*quantum mechanics*), and others, its support was too low.
- For entry word "新聞" (*newspaper*), spurious translations such as "纸张" (*paper*, *material made of cellulose pulp*), "论文" (*paper*, *scholarly article*) were successfully eliminated.
- For entry word "電車" (*electric train*), not only spurious translations such as "练" (*training*) and "教养" (*culture*) but also the correct translation "火车" were eliminated; "火车" was certainly less dominant in the Chinese corpus than another correct translation "列车."

Examples from the results of screening the Chinese-Japanese dictionary are also listed in Table 1(b).

We estimated the recall and precision of the screening results by manually assessing the correctness of translation candidates for 384 Japanese entry nouns we sampled. For comparison, we also evaluated the results of screening by using the one-time inverse consultation method (abbr. IC1) proposed by Tanaka and Umemura (1994) and its variant (abbr. IC1'). IC1 selects Chinese translation candidates linked with the Japanese entry noun via two or more English words. IC1' is the same as IC1 except that IC1' selects all translation candidates where no translation candidates are linked via two or more English words.

Table 1: Example results of screening

(a) Japanese-Chinese dictionary

Entry noun	Translation candidates	Translation candidate	Support
ホール (*hall* / *hole*)	大堂，大厅，殿，霍尔，堂，厅，洞，洞穴，洞子，空穴，孔，孔洞，孔隙，窟，窟窿，窿，漏洞，穴，阱，堀	**大厅**	**0.960**
		洞	0.034
		霍尔	0.006
新聞 (*news-paper*)	报，报纸，论文，论文儿，纸，纸头，纸张，报界，报刊，新闻界	**报纸**	**0.986**
		纸张	0.009
		论文	0.001
		纸	0.001
電車 (*electric train*)	带带，吊吊，火车，教养，练，列车，培训，培养，培育，培植，训，训练，斗车	**列车**	**0.979**
		练	0.016
		火车	0.004
		教养	0.001

(b) Chinese-Japanese dictionary

Entry noun	Translation candidates	Translation candidate	Support
大厅 (*hall*)	ホール, 会堂, 会館, 僧堂, 堂, 堂宇, 大広間, 大間, 広座敷, 広敷, 広敷き, and 5 others	**廊下**	**0.791**
		ホール	**0.203**
		会堂	0.003
		大広間	0.001
报纸 (*news-paper*)	ペイパー, ペーパ, ペーパー, ペープル, 新紙, 新聞, 新聞紙	**新聞**	**0.993**
		ペーパー	0.007
列车 (*train*)	しつけ, 仕付, 仕付け, 仕立, 仕立て, 列車, 口火, 導火, and 14 others	**列車**	**0.994**
		電車	0.006

English words.

Table 2 lists the recall and precision together with entry-noun based applicability ratios, which were calculated because IC1 produced no results for some of the entry nouns. It should be noted that the very low recall of the proposed method does not mean it is inferior. It is intended to select correct translations that are used in a corpus. However, recall is calculated under the assumption that all possible translations should be selected regardless of whether they are used or not in a corpus, since it is difficult to manually determine whether all possible translations are actually used or not in a corpus. The proposed method is superior to both IC1 and IC1' in precision. It should be added that the fairly high precision of IC1 is attained at the expense of its applicability ratio.

Table 2: Recall and precision of screening results

	Proposed method	IC1	IC1'
# entry nouns	384		
# entry nouns for which one or more translations were selected	384	273	384
Applicability ratio	100%	71.1%	100%
# correct translation candidates (S) *	2,270		
# selected translations (T)	553	926	2,485
# selected correct translations (S∩T)	359	565	1,059
Recall	15.8%	24.9%	46.7%
Precision	64.9%	61.0%	42.6%

* Correctness of 7,410 translation candidates in total for 384 entry nouns was assessed manually. Correct translations missing in the "noisy" Japanese-Chinese dictionary were not included. Therefore, recall is overestimated.

Table 3: Results of additional experiments

Method (Minimum support)	**Recall**	**Precision**
Basic method (0.1)	17.8%	71.8%
Re-screening (1st stage: 0.025, 2nd stage: 0.1)	16.8%	76.9%
Bidirectional screening – union (0.25)	26.7%	66.8%
Bidirectional screening – intersection (0.025)	18.0%	76.2%

Note: Results for 247 Japanese entry nouns were evaluated.

3.3 Additional Experiments

3.3.1 Re-screening

Correlation matrices calculated by using a noisy bilingual dictionary, which increases ambiguity in the alignment of word associations, may be less reliable than those calculated by using a noiseless bilingual dictionary. Therefore, we evaluated the method of re-screening in which a bilingual dictionary is screened again by using correlation matrices recalculated with the bilingual dictionary once screened. Considering the purpose of first-stage screening, the minimum support at the first stage was set to a very low value of 0.025, while it was set to 0.1 at the second stage. The recall and precision of the re-screening results for 247 Japanese entry nouns we sampled are listed in Table 3; re-screening improved precision by about 5% while the recall was comparable to that with the basic method.

3.3.2 Combined bidirectional screening

Although bilingual dictionaries are reversible, the results from screening their reverse do not always coincide with the reversed results from screening these same dictionaries. Therefore, proper combination of the screening results of a Japanese-Chinese dictionary with those of a Chinese-Japanese dictionary can produce better results.

(1) Union of bidirectional screening results

The proposed method has a shortcoming in that, among synonymous translation candidates, those except the most dominant one are underestimated because most of their associated words are taken by the most dominant one; for example, in Table 1, the support for "火车" was very low because most of its associated words were taken by "列车." This may be alleviated by screening the reverse dictionary since the proposed method works well even for less dominant entry words. Therefore, we evaluated the union of a screened Japanese-Chinese dictionary and the reverse of a screened Chinese-Japanese dictionary, where the minimum support was set to a rather high value, i.e., 0.25, since less dominant translations were expected to be in the results of screening in the opposite direction. The recall and precision of the union of bidirectional screening results are listed in Table 3; compared to the basic method, recall has improved significantly although precision has decreased by about 5%.

(2) Intersection of bidirectional screening results

A pair of Japanese and Chinese words resulting from screening both Japanese-Chinese and Chinese-Japanese dictionaries may be reliable even if its support is not high in either direction. Therefore, we evaluated the intersection of a screened Japanese-Chinese dictionary and the reverse of a screened Chinese-Japanese dictionary, where the minimum support was set to a rather low value, i.e., 0.025. The recall and precision of the intersection of bidirectional screening results are also listed in Table 3; compared to the basic method, precision has improved by about 4% while recall has remained almost the same.

4. Discussion

Although the experimental results demonstrate that the proposed method is viable, there is much room for improvement. It is crucial to optimize parameters including the window size, the word frequency threshold, the mutual information threshold, the number of associated words per entry word, and the minimum support. The experiments described in Section 3 were carried out with rather arbitrary parameter settings. We found that the parameters in extracting word associations predominantly affect the screening results. The parameters in calculating the correlation matrices of associated words versus translations, i.e., the relative weight between the primary and secondary correlation factors and the number of iterations, are less problematic; Kaji and Morimoto (2005) confirmed that the iterative algorithm works stably within a rather wide range of parameter values.

The additional experiments described in Subsection 3.3 demonstrate the effectiveness of both re-screening and bidirectional screening. Re-screening can of course be combined with bidirectional screening. We should choose unions or intersections in bidirectional screening depending on which is preferable, high recall or high precision.

Obviously, the resulting bilingual dictionary not only depends on parameter settings but also source bilingual dictionaries; the proposed method cannot select any translations if they are not in a combination of source dictionaries. We did another additional experiment in which the same Japanese-Chinese noun dictionary as in Section 3 was screened after being manually supplemented with appropriate translation candidates. Table 4 compares the results of screening the translation candidates of three Japanese entry nouns before and after supplementation. We can expect that supplementation with appropriate translation candidates will certainly improve not only recall but also precision. Thus, we should use much higher-coverage dictionaries than the EDR Japanese-English Bilingual Dictionary and the LDC Chinese-English Translation Lexicon from a practical point of view.

5. Related work

Methods of automatically constructing a bilingual dictionary via a third language are divided into those producing generic dictionaries and those producing domain-dependent dictionaries. The former include Tanaka and Umemura's (1994) one-time/two-time inverse consultation method and its derivatives (Bond, et al., 2001; Shirai and Yamamoto, 2001; Zhang, et al., 2007). The latter are few as far as work explicitly addressing the problem is concerned. However, methods of acquiring bilingual lexicons from nonparallel corpora based on contextual similarities (Rapp, 1995; Kaji and Aizono, 1996; Tanaka and Iwasaki, 1996; Fung and Yee, 1998; Rapp, 1999) can obviously be applied to the problem. Recently, Sammer and Soderland (2007) proposed a method using contextual similarities to construct a multilingual lexicon, called PanLexicon, from monolingual corpora and bilingual lexicons.

Table 4: Example results of screening manually supplemented dictionary

Entry noun	**Before supplementation**	**After supplementation**
難民 (*refugees*)	受难者	难民
先端 (*point*, *tip*)	终端，小费，论点	顶端
故郷 (*hometown*)	底座，由来	家乡，故乡

Methods based on contextual similarities and the proposed method share an underlying assumption that the translations of words that are associated in one language will also be associated in the other language. The difference between them originates from the difference in their original purpose, i.e., bilingual lexicon acquisition versus word sense disambiguation. As a method for eliminating spurious translations, the proposed method has an advantage in that it is easy to set a minimum support common to all entry words while it is difficult to set a contextual similarity threshold common to all entry words. The proposed method also has a distinct feature in that the resulting bilingual dictionary, which provides translations together with associated words supporting all translations, enables translations to be selected according to contexts.

6. Conclusion

We developed a method of constructing a "domain-adapted" bilingual dictionary for a new language pair from two bilingual dictionaries that share one of the languages. The method requires the monolingual corpora of the respective languages, whose availability is not as limited as that of parallel corpora. It correlates the associated words of an entry word with its translation candidates resulting from a combination of the source bilingual dictionaries, and it then selects translation candidates that have the highest correlations with a certain percentage or more of the associated words. The main feature of the proposed method is that the resulting bilingual dictionary, which not only provides translations but also associated words supporting all translations, enables contextually based selection of translations. Preliminary experiments using the EDR Japanese-English and LDC Chinese-English dictionaries and Mainichi Newspaper and Xinhua News Agency corpora demonstrate that the proposed method is viable. A major problem that remains is optimization of parameters.

7. Acknowledgements

This work was supported by the Japanese Government's Special Coordination Funds for Promoting Science and Technology and Hitachi, Ltd.

8. 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 Machine Translation Summit VIII*, pp. 53-58.

Church, Kenneth and Patrick Hanks. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1): 22-29.

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 of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pp. 414-420.

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*, pp. 23-28.

Kaji, Hiroyuki and Yasutsugu Morimoto. (2002). Unsupervised word sense disambiguation using bilingual comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 411- 417.

Kaji, Hiroyuki and Yasutsugu Morimoto. (2005). Unsupervised word-sense disambiguation using bilingual comparable corpora. *IEICE Transactions on Information and Systems*, E88-D(2): 289-301.

Nakagawa, Tetsuji and Kiyotaka Uchimoto. (2007). A hybrid approach to word segmentation and POS tagging, In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pp. 217-220.

Rapp, Reinhard. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 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*, pp. 519-526.

Sammer, Marcus and Stephen Soderland. (2007). Building a sense-distinguished multilingual lexicon from monolingual corpora and bilingual lexicons. In *Proceedings of Machine Translation Summit XI*, pp. 399-406.

Shirai, Satoshi and Kazuhide Yamamoto. (2001). Linking English words in two bilingual dictionaries to generate another language pair dictionary. In *Proceedings of the 19th International Conference on Computer Processing of Oriental Languages*, pp. 174-179.

Tanaka, Kumiko and Hideya Iwasaki. (1996). Extraction of lexical translations from non-aligned corpora. In *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 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*, pp. 297-303.

Zhang, Yujie, Qing Ma, and Hitoshi Isahara. (2007). Building Japanese-Chinese translation dictionary based on EDR Japanese-English bilingual dictionary. In *Proceedings of Machine Translation Summit XI*, pp. 551-557.

Appendix: Algorithm for correlating associated words with word senses

The first language has been assumed to be English and the second Japanese. Bilingual comparable corpora, unlike parallel corpora, cannot be used as training data for supervised word sense disambiguation since they do not contain word-for-word alignment. However, it is possible to extract word associations from the corpora of both languages and align them across languages with the assistance of a bilingual dictionary; e.g., a word association (tank, soldier) extracted from an English corpus would be aligned with its counterpart (戦車, 兵士) extracted from a Japanese corpus. Note that aligned word-associations reveal the sense the associated word of a polysemous word supports; e.g., the alignment of (tank, soldier) with (戦車, 兵士) reveals that the associated word "soldier" of the polysemous word "tank" supports the military vehicle sense of "tank." Thus, a pair of bilingual comparable corpora enables unsupervised word sense disambiguation assuming that each word sense is defined as a set of synonymous translations (Kaji and Morimoto, 2002).

This naive idea, however, is hampered by alignment failure as well as ambiguity in alignment. For example, the word association (tank, ozone) extracted from an English corpus could not be aligned with its counterpart (タンク, オゾン) unless its counterpart was extracted from a Japanese corpus, and, therefore, which sense of "tank" the associated word "ozone" supported could not be determined. In addition, another English word association (tank, troop) could be aligned with Japanese word associations such as (戦車 (*military vehicle*), 隊 (*a body of soldier*)) and (水槽 (*water tank*), 群れ (*a band of animals*)), and, therefore, which sense of "tank" the associated word "troop" supported could not be determined.

To cope with alignment failure, the primary correlation factor is defined based on the assumption that associated words that are associated with one another support the same sense. For example, the following is a set consisting of words that are associated not only with "tank" but also with "ozone."

A(tank, ozone) = {air, area, car, control, deep, defense, emission, fuel, gas, gasoline, pump, road, study, upper, vapor}.

This suggests that the associated word "ozone" should support the same sense of "tank" as (most of) the associated words such as "air," "area," "car," and "control." Thus, the primary correlation factor of "ozone" with a sense of "tank" should be proportional to the sum of correlations of "air," "area," "car," "control," and others with the same sense. That is,

$$C'(\text{ozone}, \{\text{戦車}\}) \propto \sum_{x \in A(\text{tank},\text{ozone})} C(x, \{\text{戦車}\}).$$

$$C'(\text{ozone}, \{\text{タンク}, \text{水槽}, \text{槽}\}) \propto \sum_{x \in A(\text{tank},\text{ozone})} C(x, \{\text{タンク}, \text{水槽}, \text{槽}\}).$$

Note that the military vehicle sense of "tank" is defined as {戦車} and its container sense is defined as {タンク, 水槽, 槽}.

To cope with ambiguity in alignment, the secondary correlation factor is defined based on the following assumption: since the alignment of "triplets" of words that are associated with one another is much more reliable than the alignment of "pairs" of words that are associated with one another, the plausibility of an alignment of word associations, i.e., pairs of words, could be estimated by how many alignments of triplets contain the alignment of word associations in question. For example, the following is a set consisting of words each of which, together with its Japanese translation, makes up an alignment of triplets containing the alignment of (tank, troop) with (戦車, 隊).

B((tank, troop), (戦車, 隊)) = {air, area, army, battle, commander, defense, fight, fire, force, government, helicopter, Russia, Serb, soldier}.

Likewise, the following is a set consisting of words each of which, together with its Japanese translation, makes up an alignment of triplets containing the alignment of (tank, troop) with (水槽,群れ).

B((tank, troop), (水槽, 群れ)) = {air, area, fire, government}.

Comparing these sets of associated words reveals that the alignment of (tank, troop) with (戦車, 隊) is more plausible than the alignment of (tank, troop) with (水槽, 群れ). The secondary correlation factor of the associated word "troop" with the military vehicle sense of "tank" should be proportional to the sum of correlations of "air," "area," "army," "battle," and others with the military vehicle sense, and that with the container sense of "tank" should be proportional to the sum of correlations of "air," "area," "fire," and "government" with the container sense. That is,

$$C''(\text{troop}, \{\text{戦車}\}) \propto \max_{y \in \{\text{戦車}\}, y'} \sum_{x \in B((\text{tank},\text{troop}),(y,y'))} C(x, \{\text{戦車}\}).$$

$$C''(\text{troop}, \{\text{タンク}, \text{水槽}, \text{槽}\}) \propto \max_{y \in \{\text{タンク},\text{水槽},\text{槽}\}, y'} \sum_{x \in B((\text{tank},\text{troop}),(y,y'))} C(x, \{\text{タンク}, \text{水槽}, \text{槽}\}).$$

It is naturally assumed that the mutual information between two words reflects the correlation between their relevant senses. Consequently, the correlation between the *i*-th associated word $x'(i)$ of a word, x, and its j-th sense $s(x, j)$ is defined by the following formulas.

$$C\big(x'(i), s(x, j)\big) = MI\big(x'(i), x\big) \cdot \frac{C'\big(x'(i), s(x, j)\big) + \alpha \cdot C''\big(x'(i), s(x, j)\big)}{\max_k \big[C'\big(x'(i), s(x,k)\big) + \alpha \cdot C''\big(x'(i), s(x,k)\big)\big]}, \quad [2']$$

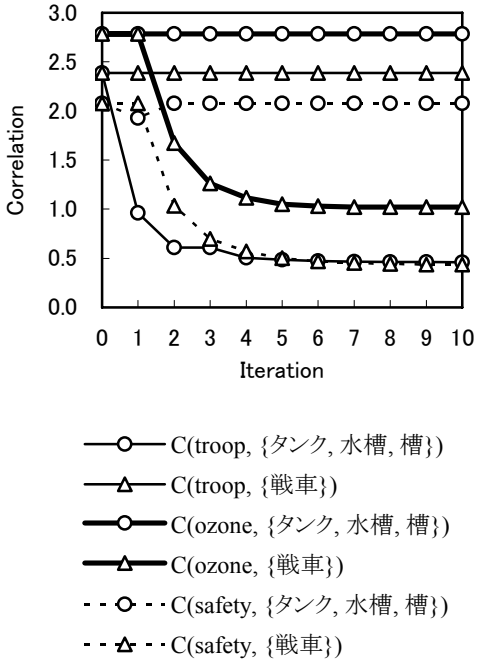


Fig. 4: Convergence of correlations

where α is a parameter specifying the relative weight between primary and secondary correlation factors C' and C''.

$$C'\big(x'(i), s(x, j)\big) = \sum_{x'' \in A(x, x'(i))} C\big(x'', s(x, j)\big), \quad [3']$$

where $A(x, x'(i))$ denotes a set consisting of words each of which is associated with both x and $x'(i)$.

$$C''\big(x'(i), s(x, j)\big) = \max_{y \in s(x, j), y'} \left(MI(y', y) \cdot \sum_{x'' \in B((x, x'(i)),(y, y'))} C\big(x'', s(x, j)\big) \right), \quad [4']$$

where $B((x, x'(i)), (y, y'))$ denotes a set consisting of words each of which is associated with both x and $x'(i)$ and has at least one translation associated with both y and y' where word association $(x, x'(i))$ can be aligned with word association (y, y').

Note that the above definition of correlations between associated words and senses is recursive. The correlations can be calculated iteratively with the following initial values:

$$C\big(x'(i), s(x, j)\big) = MI\big(x'(i), x\big). \quad [5']$$

We confirmed through experiments that the correlation values converged within 10 iterations, as seen in Fig. 4. It should be added that formulas [2] to [5] in Section 2.3 is a special case of formulas [2′] to [5′] in which each translation $y(j)$ of x defines a distinct sense of x, i.e., $s(x, j) = \{ y(j) \}$.