Building Bilingual Lexicons Using Lexical Translation Probabilities via Pivot Languages

Takashi Tsunakawa[†], **Naoaki Okazaki**[†], **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 / National Centre for Text Mining
131 Princess Street, Manchester, M1 7DN, UK
{tuna, okazaki, tsujii} at is.s.u-tokyo.ac.jp

Abstract

This paper proposes a method of increasing the size of a bilingual lexicon obtained from two other bilingual lexicons via a pivot language. When we apply this approach, there are two main challenges, *ambiguity* and *mismatch* of terms; we target the latter problem by improving the utilization ratio of the bilingual lexicons. Given two bilingual lexicons between language pairs L_f–L_p and L_p–L_e, we compute lexical translation probabilities of word pairs by using a statistical word-alignment model, and term decomposition/composition techniques. We compare three approaches to generate the bilingual lexicon: *exact merging*, *word-based merging*, and our proposed *alignment-based merging*. In our method, we combine lexical translation probabilities and a simple language model for estimating the probabilities of translation pairs. The experimental results show that our method could drastically improve the number of translation terms compared to the two methods mentioned above. Additionally, we evaluated and discussed the quality of the translation outputs.

1. Introduction

Bilingual lexicon is a crucial resource for cross-lingual applications of natural language processing (NLP) including machine translation (Brown et al., 1990), and cross-lingual information retrieval (Nie et al., 1999). Thus, a number of bilingual lexicons were constructed despite its expensive compilation costs. However, it is unrealistic to construct a bilingual lexicon for every language pair; the number of language pairs would be as many as 4,950, given that there were 100 languages in the world. Moreover, it is difficult to maintain a bilingual lexicon with the rapid growth of neologism. Consequently, comprehensible bilingual lexicons are available only for small subsets of language pairs, and are unavailable for most language pairs.

To address this problem, researchers have proposed the use of pivot languages (third languages) as an intermediary language to construct bilingual lexicons automatically (Tanaka and Umemura, 1994; Bond et al., 2001; Shirai and Yamamoto, 2001; Paik et al., 2001; Schafer and Yarowsky, 2002; Zhang et al., 2005; Goh et al., 2005), and recently, a commercial machine translation system[1] implemented the pivot approach for automatically searching phrase or sentence pairs. The basic idea of this approach is to create a bilingual lexicon between two languages L_e and L_f, by merging two large bilingual lexicons, L_e–L_p and L_p–L_f, where L_p is the pivot language. The advantage to this approach is that we can obtain a bilingual lexicon between L_e and L_f, even if no bilingual lexicon exists between these languages. However, the approach also presents two major challenges; these are *ambiguity* and *mismatch*.

In general, it is not guaranteed that the word w_e (in language L_e), translated from a word w_f (in language L_f) via a pivot word w_p (in language L_p), is correct, especially when the pivot word w_p is polysemous. For example, a Japanese term "土手," *dote*: embankment, levee, may be associated with a Chinese term "银行," *yínghán*g: banking institution, finance institution, using the pivot word "bank" in English. In order to solve the ambiguity problem in pivot terms, Tanaka et al. (1994) proposed the use of the structure of bilingual dictionaries to select correct translation equivalents. Bond et al. (2001) utilized semantic classes to rank translation equivalents; word pairs with compatible semantic classes are preferred to those with dissimilar classes. Shirai et al. (2001) measured the number of words in a pivot language shared by a translation pair to measure the similarity of the two words in the target languages. Paik et al. (2001) used multiple pivot languages (English and Chinese) to improve the accuracy of dictionary construction. Schafer et al. (2002) presented a method for inducing translation lexicons between two distant languages via a bridge language, using cross-language context similarity, weighted Levenshtein distance, relative frequency, and burstiness similarity measures.

Another issue arises in merging terms in the pivot language L_p from different bilingual lexicons. Since two bilingual lexicons L_f–L_p, and L_p–L_e are constructed independently, we cannot assume that the two lexicons use the identical term to describe a single entity. For example, it is impossible to associate two translation pairs ("地球温暖化 (*chikyū-ondanka*)," "global warming"), and ("global heating," "全球变暖 (*quánqiú-biànnuǎn*)") because of the different terms in the pivot language. In addition, bilingual lexicons developed for technical terms may contain a number of terms that cannot be associated with other lexicons. For example, even if a Japanese–English lexicon is large enough to include a technical term, "石炭転換プロセス (*sekitan-tenkan-purosesu*)" (coal conversion process), we can obtain its Chinese translation, "煤转化过程 (*méizhuǎnhuà-guòchéng*)" only when the Chinese-English lexicon includes the English term as it is.

This paper presents a solution to the latter problem, that is, to increase the number of translation pairs obtained from

[1] http://www.esteam.se

two bilingual lexicons, assuming that the former problem should be dealt with within the succeeding step. Given two large bilingual lexicons L_f–L_p, and L_p–L_e, we compute the translation probability from a word, w_f, to w_e by using a statistical word-alignment model, and term decomposition/composition techniques. After collecting term pairs, the evaluation of the correctness of translations, an intelligent suggestion system for dictionary editors, etc. might be necessary for constructing a more sophisticated system. These topics are beyond the scope of this paper.

2. Merging Two Bilingual Lexicons

Let L_f, L_p, and L_e be monolingual lexicons in source, pivot, and target languages, respectively. Suppose that we have two bilingual lexicons L_f–L_p and L_p–L_e:

$$L_f\text{–}L_p = \{(\bar{w}_f, \bar{w}_p) | \bar{w}_f \text{ is a translation of } \bar{w}_p\} \quad (1)$$

$$L_p\text{–}L_e = \{(\bar{w}_p, \bar{w}_e) | \bar{w}_p \text{ is a translation of } \bar{w}_e\}, \quad (2)$$

where $\bar{w}_f, \bar{w}_p$, and $\bar{w}_e$ denotes the terms in the lexicons L_f, L_p, and L_e respectively.

The simplest method for constructing the L_f–L_e lexicon is to connect source and target terms that share a common translation term in the pivot language:

$$\begin{aligned} L_f\text{–}{L_e}^{(e)} &= \{(\bar{w}_f, \bar{w}_e) | \exists \bar{w}_p \, ((\bar{w}_f, \bar{w}_p) \in L_f\text{–}L_p \\ &\quad \wedge (\bar{w}_p, \bar{w}_e) \in L_p\text{–}L_e)\}. \end{aligned} \quad (3)$$

We call this algorithm *exact merging*.

It is a straightforward extension to decompose a source term into a sequence of constituent words, and to consult the lexicon built by the above method in order to translate the words in the source term into target words one by one. That is,

$$\begin{aligned} L_f\text{–}{L_e}^{(w)} = \{(\bar{w}_f, \bar{w}_e) | \forall i = 1, \ldots, l \\ ((w_{fi}, w_{ei}) \in L_f\text{–}L_e^{(e)}\} \cup L_f\text{–}{L_e}^{(e)}, \end{aligned} \quad (4)$$

where $w_{f1}, \ldots, w_{fl}$ and $w_{e1}, \ldots, w_{fl}$ are sequences of constituent words of $\bar{w}_f$ and $\bar{w}_e$, respectively. We call this algorithm *word-based merging*.

However, the constituent words of source terms are not always included in the lexicon L_f–${L_e}^{(e)}$. In addition, neither exact merging nor word-based merging provides a confidence value that indicates that two words are translation equivalents, useful for machine translation systems.

Recently, several researchers proposed the use of the pivot language for phrase-based statistical machine translation (Utiyama and Isahara, 2007; Wu and Wang, 2007). In these approaches, the translation probabilities between source and target terms are calculated via the pivot terms. Similarly, we introduce a statistical word-alignment model for estimating the translation probabilities between source and target words. We calculate the term translation probabilities by using the product of translation probabilities of constituent words.

We obtain word alignments $a_{e\text{–}p}$ and $a_{p\text{–}f}$ of the lexicons L_e–L_p and L_p–L_f by GIZA++ and the refinement method (Och and Ney, 2003). The lexical translation probabilities are calculated as follows:

$$p(w_p | w_e; a_{e\text{–}p}) = \frac{C(w_e, w_p, a_{e\text{–}p})}{C(w_e)}, \quad (5)$$

$$p(w_f | w_p; a_{p\text{–}f}) = \frac{C(w_p, w_f, a_{p\text{–}f})}{C(w_p)}, \quad (6)$$

$$\begin{aligned} & p(w_f | w_e; a_{e\text{–}p}, a_{p\text{–}f}) \\ = & \sum_{w_p \in L_p} p(w_f | w_p; a_{p\text{–}f}) p(w_p | w_e; a_{e\text{–}p}). \end{aligned} \quad (7)$$

In these equations, $C(w_e)$ denote the frequency of the word w_e in the lexicon L_e–L_p, $C(w_p)$, the frequency of the word w_p in the lexicon L_p–L_f, and $C(w_e, w_p, a_{e\text{–}p})$, and the co-occurrence frequency of w_e and w_p when they are aligned by $a_{e\text{–}p}$.

Equation 8 computes the translation probability from $\bar{w}_e$ to $\bar{w}_f$,

$$p(\bar{w}_f | \bar{w}_e; a_{e\text{–}p}, a_{p\text{–}f}) = \prod_{i=1}^{l} p(w_{fi} | w_{ei}; a_{e\text{–}p}, a_{p\text{–}f}). \quad (8)$$

Finally, we obtain the probability of $p(\bar{w}_e | \bar{w}_f)$ by using the noisy-channel model:

$$\begin{aligned} p(\bar{w}_e | \bar{w}_f) &= \frac{p(\bar{w}_f | \bar{w}_e; a_{e\text{–}p}, a_{p\text{–}f}) p(\bar{w}_e)}{p(\bar{w}_f)} \\ &\propto p(\bar{w}_f | \bar{w}_e; a_{e\text{–}p}, a_{p\text{–}f}) p(\bar{w}_e). \end{aligned} \quad (9)$$

In order to estimate the monolingual language model $p(\bar{w}_e)$, we use the Google[2] hit count (the number of retrieved pages) by querying the term $\bar{w}_e$. Assuming that the total number of Web pages is a constant N, we estimate the probability $p(\bar{w}_e)$,

$$p(\bar{w}_e) = \frac{\text{(hit count of } \bar{w}_e)}{N}. \quad (10)$$

We can thus generate the merged lexicon with translation probabilities by using:

$$\begin{aligned} L_f\text{–}{L_e}^{(a)} = \{(\bar{w}_f, \bar{w}_e, p(\bar{w}_e | \bar{w}_f), p(\bar{w}_f | \bar{w}_e) | \\ p(\bar{w}_e | \bar{w}_f) > 0 \wedge p(\bar{w}_f | \bar{w}_e) > 0\}. \end{aligned} \quad (11)$$

We call this algorithm *alignment-based* merging.

3. Experiment

3.1. Data

We used Japanese-English and English-Chinese lexicons to build a Japanese-Chinese lexicon. The Japanese-English lexicon, which was released by the Japan Science and Technology Agency (JST)[3], consists of 527,206 translation equivalents (465,572 Japanese terms and 418,044 English terms) extracted from academic papers on science and technology. It covers a wide range of named entities such as company, place, and chemical names that may be difficult to translate into English and Japanese terms. The Chinese-English lexicon, which was compiled by Wanfang Data Co., Ltd[4], includes 525,259 translation equivalents (441,710 Chinese terms and 430,501 English terms) in the field of scientific research.

Lexicon	**# of L_J**	**# of L_E**	**# of L_C**
L_J–L_E	465,543	416,578	-
L_E–L_C	-	429,766	439,795
L_E	-	777,344	-
L_J–${L_C}^{(e)}$	103,437 (22.2%)	68,996	98,537 (22.4%)
L_J–${L_C}^{(w)}$	124,945 (26.8%)	-	167,929 (38.1%)
L_J–${L_C}^{(a)}$	438,976 (94.2%)	-	342,229 (77.8%)

Table 1: The statistics of merged lexicons

3.2. Size of Merged Lexicon

We generated three lexicons merged by exact, word-based, and alignment-based methods. All terms in Japanese-English and Chinese-English lexicons were lower-cased in advance. We employed the following word tokenizers: JUMAN[5] for Japanese, a Maximum Entropy Markov Model (MEMM)-based part-of-speech tagger[6] (Tsuruoka and Tsujii, 2005) for English, and the morphological tokenizer "cjma" (Nakagawa and Uchimoto, 2007) for Chinese. Table 1 shows the distinct numbers of terms in the original and merged lexicons, and the *utilization ratio* in parentheses (the number of terms in the original lexicon used for building the merged lexicon).

The exact merging translated 103,437 (22.2%) of Japanese terms into Chinese, and 98,537 (22.4%) of Chinese terms into Japanese. These figures imply that about 80% of the terms remained unused in building the Japanese-Chinese lexicon. The word-based merging translated 124,945 (26.8%) of Japanese terms and 167,929 (38.1%) of Chinese terms; this brings 4.62% of the Japanese terms and 15.8% Chinese terms into the bilingual lexicon. In contrast, the alignment-based merging constructed a Japanese-Chinese bilingual lexicon with 438,976 (94.2%) Japanese terms and 342,229 (77.8%) Chinese terms. The utilization ratio was drastically improved from the exact method, and the size of the merged bilingual lexicon also increased.

3.3. Accuracy of Merged Lexicon

We evaluated the accuracy of the bilingual translation pairs obtained by the proposed method. 50 Japanese and 50 Chinese evaluation terms were chosen at random from a set of terms that were not translated into another language by the word-based method. Obtaining the top-10 translation equivalents with high scores for each evaluation term, we asked two human subjects[7] who are fluent in both Japanese and Chinese to judge the correctness of the translation equivalents.

We employed the precision and mean reciprocal rank (MRR) (Voorhees, 1999). We define the precision as the ratio of source terms that are successfully mapped to its translation only if one of ten translation equivalents includes the correct translation. MRR is calculated as follows. We sort the translation equivalents for each source term $\bar{w}_f$ by the probability $p(\bar{w}_e|\bar{w}_f)$. Each source term $\bar{w}_f$ receives a score equal to the reciprocal of the rank at which the first correct translation $\bar{w}_e$ is obtained. After that, we calculate the mean of reciprocal ranks over all source terms.

Table 2 shows the MRR scores and the precisions. "Prec1" is the precision of the highest ranked terms, and "Prec10" is the precision that the 10-best outputs include the correct

Source	Target	MRR	Prec1	Prec10
Japanese	Chinese	0.242	0.14	0.46
Chinese	Japanese	0.258	0.20	0.40

Table 2: Mean reciprocal scores and precisions

Japanese	English	Score	P	H
角膜 实質炎 [T]	kerato-parenchymatitis	0.557	-2.89	432
角膜 的 炎	kerato- inflammation	0.00457	-3.34	10
角膜 物質炎	kerato- material inflammation	0	-2.24	0
角膜 物質関節	kerato- material joint	0	-2.49	0
角膜 実 炎	kerato- real inflammation	0	-2.63	0
角膜 物質性	kerato- materiality	0	-2.66	0
角膜 材料炎	kerato- stuff inflammation	0	-2.66	0
角膜 物質高安	kerato- material *Takayasu*	0	-2.83	0
角膜 物質胃腸	kerato- material stomach	0	-2.87	0

Table 3: An example of translation of "角膜 实质炎" (keratitis parenchymatosa) according to alignment-based merging: [T] is the correct translation, $P = \log_{10} p(\bar{w}_f|\bar{w}_e; a_{e-p}, a_{p-f})$, H = (hit count), and Score = $p \times H$.

Chinese	English	Score	P	H
的 状态	state of	7249	-2.43	1960000
发展 状态	development state	6593	-1.58	252000
发展 条件	development condition	6001	-2.05	674000
的 条件	condition of	3159	-2.90	2510000
发展 国家	development country	2715	-2.57	998000
生长 状态 [T]	growing state	2688	-1.51	87900
生长 条件	growing condition	2248	-1.98	216000
增长 状态 [T]	rising state	1343	-1.72	69800
开发 条件	development condition	1260	-2.18	192000

Table 4: An example of translation of "発育 状態" (growth status) according to alignment-based merging

[2] http://www.google.com/

[3] http://pr.jst.go.jp/others/tape.html

[4] http://www.wanfangdata.com/

[5] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[6] http://www-tsujii.is.s.u-tokyo.ac.jp/ tsuruoka/postagger/

[7] One subject was for Japanese-to-Chinese and another for Chinese-to-Japanese.

one. The proposed method generated correct translations for half of terms that could not be associated by the word-based merging. The MRR score indicated that the proposed method ranked the correct translations at the 4th place on average.

Tables 3 and 4 illustrate examples of translation pairs obtained by the proposed method. In Table 3, the correct translation for the source term, "角膜 实质 炎," (keratitis parenchymatosa) appeared on the top. In contrast, the correct translation could not appear on higher ranks but ranked 6th and 8th in Table 4. This is because incorrect translations are used frequently in Chinese to represent other senses.

There were several kinds of errors in the outputs, and the most frequent errors are caused by inappropriate tokenization, and errors from data sparseness. For example, a Chinese input term "大孢 子吸器 (megaspore haustorium)" should be tokenized into "大孢子 (megaspore)," and "吸器 (haustorium)," for finding the correct translation. Similarly, the tokenizer could not split "ターンシグナルフラッシャ (turn signal flasher)" into "ターン (turn)," "シグナル (signal)," and "フラッシャ (flasher)," and the system could not find appropriate word alignments. This problem could be solved by improving the accuracy of the tokenizers, and introducing phrase-based model for machine translation.

4. Conclusion

This paper presented an approach to increase the number of translation pairs obtained from two bilingual lexicons via a pivot language. The experimental results confirmed that the proposed method improves the utilization ratio of the existing bilingual lexicons drastically. The proposed method does not include a mechanism to improve the precision, e.g., to choose a correct translation by examining the context or semantic classes of source and target terms. A future direction of this study would be to combine more sophisticated scoring methods for translation equivalents to improve the precision of the merged bilingual lexicon. We are also planning on evaluating a machine translation system with this lexicon integrated to confirm the contribution of the bilingual lexicon.

5. Acknowledgements

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 the Japan Science and Technology Agency (JST) for providing a useful bilingual lexicon.

6. References

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

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsumoto. 2005. Building a Japanese-Chinese dictionary using kanji/hanzi conversion. In *Proc. of the 2nd International Joint Conference on Natural Language Processing*, pages 670–681.

Tetsuji Nakagawa and Kiyotaka Uchimoto. 2007. Hybrid approach to word segmentation and POS tagging. In *Companion Volume to the Proc. of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 217–220.

Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–81.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kyonghee Paik, Francis Bond, and Shirai Satoshi. 2001. Using multiple pivots to align Korean and Japanese lexical resources. In *Proc. of the Workshop on Language Resources in Asia, Natural Language Processing Pacific Rim Symposium 2001*, pages 63–70.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proc. of the 6th Conference on Natural Language Learning*, volume 20, pages 1–7.

Satoshi Shirai and Kazuhide Yamamoto. 2001. Linking English words in two bilingual dictionaries to generate another language pair dictionary. In *Proc. of 19th International Conference on Computer Processing of Oriental Language*, pages 174–179.

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *Proc. of the 15th International Conference on Computational Linguistics*, pages 297–303.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proc. of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 467–474.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proc. of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 484–491.

Ellen M. Voorhees. 1999. The TREC-8 question answering track report. In *Proc. of TREC-8*, pages 77–82.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 856–863.

Yujie Zhang, 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.