Building a Bilingual Lexicon Using Phrase-based Statistical Machine Translation via a Pivot Language

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}@is.s.u-tokyo.ac.jp

Abstract

This paper proposes a novel method for building a bilingual lexicon through a pivot language by using phrase-based statistical machine translation (SMT). Given two bilingual lexicons between language pairs L_f–L_p and L_p–L_e, we assume these lexicons as parallel corpora. Then, we merge the extracted two phrase tables into one phrase table between L_f and L_e. Finally, we construct a phrase-based SMT system for translating the terms in the lexicon L_f–L_p into terms of L_e and, obtain a new lexicon L_f–L_e. In our experiments with Chinese-English and Japanese-English lexicons, our system could cover 72.8% of Chinese terms and drastically improve the utilization ratio.

1 Introduction

The bilingual lexicon is a crucial resource for multilingual applications in natural language processing including machine translation (Brown et al., 1990) and cross-lingual information retrieval (Nie et al., 1999). A number of bilingual lexicons have been constructed manually, despite their expensive compilation costs. However, it is unrealistic to build a bilingual lexicon for every language pair; thus, comprehensible bilingual lexicons are available only for a limited number of language pairs.

One of the solutions is to build a bilingual lexicon of the source language L_f and the target L_e through a pivot language L_p, when large bilingual lexicons L_f–L_p and L_p–L_e are available. Numerous researchers have explored the use of pivot languages (Tanaka and Umemura, 1994; Schafer and Yarowsky, 2002; Zhang et al., 2005). This approach is advantageous because we can obtain a bilingual lexicon between L_e and L_f, even if no bilingual lexicon exists between these languages.

Pivot-based methods for dictionary construction may produce incorrect translations when the word w_e is translated from a word w_f by a polysemous pivot word w_p[1]. Previous work addressed the polysemy problem in pivot-based methods (Tanaka and Umemura, 1994; Schafer and Yarowsky, 2002). Pivot-based methods also suffer from a mismatch problem, in which a pivot word w_p from a source word w_f does not exist in the bilingual lexicon L_p–L_e[2]. Moreover, a bilingual lexicon for technical terms is prone to include a number of pivot terms that are not included in another lexicon.

This paper proposes a method for building a bilingual lexicon through a pivot language by using phrase-based statistical machine translation (SMT) (Koehn et al., 2003). We build a translation model between L_f and L_e by assuming two lexicons L_f–L_p and L_p–L_e as parallel corpora, in order to increase the obtained lexicon size by handling multi-word expressions appropriately. The main advantage of this method is its ability to incorporate various translation models that associate languages L_f–L_e; for example, we can further improve the translation model by integrating a small bilingual lexicon L_f–L_e.

[1] A Japanese term "土手": *dote*, embankment, may be associated with a Chinese term "银行," *yíngháng*: banking institution, using the pivot word *bank* in English.

[2] 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 difference in English (pivot) terms.

Coling 2008: Companion volume – Posters and Demonstrations, pages 127–130
Manchester, August 2008

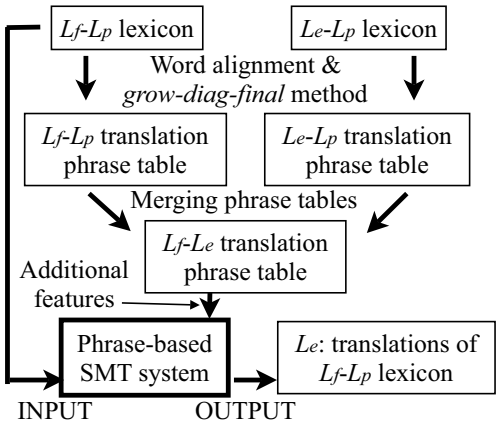

Figure 1: Framework of our approach

2 Merging two bilingual lexicons

We introduce phrase-based SMT for merging the lexicons, in order to improve both the merged lexicon size and its accuracy. Recently, several researchers proposed the use of the pivot language for phrase-based SMT (Utiyama and Isahara, 2007; Wu and Wang, 2007). We employ a similar approach for obtaining phrase translations with the translation probabilities by assuming the bilingual lexicons as parallel corpora. Figure 1 illustrates the framework of our approach.

Let us suppose that we have two bilingual lexicons L_f–L_p and L_p–L_e. We obtain word alignments of these lexicons by applying GIZA++ (Och and Ney, 2003), and *grow-diag-final* heuristics (Koehn et al., 2007). Let $\bar{w}_x$ be a phrase that represents a sequence of words in the language L_x. For phrase pairs $(\bar{w}_p, \bar{w}_f)$ and $(\bar{w}_e, \bar{w}_p)$, the translation probabilities $p(\bar{w}_p|\bar{w}_f)$ and $p(\bar{w}_e|\bar{w}_p)$ are computed using the maximum likelihood estimation from the co-occurrence frequencies, consistent with the word alignment in the bilingual lexicons. We calculate the direct translation probabilities between source and target phrases,

$$p(\bar{w}_e|\bar{w}_f) = \frac{\sum_{\bar{w}_p} p(\bar{w}_e|\bar{w}_p)p(\bar{w}_p|\bar{w}_f)}{\sum_{\bar{w}'_e}\sum_{\bar{w}_p} p(\bar{w}'_e|\bar{w}_p)p(\bar{w}_p|\bar{w}_f)}. \quad (1)$$

We employ the log-linear model of phrase-based SMT (Och and Ney, 2002) for translating the source term $\bar{w}_f$ in the lexicon L_f–L_p into the target language by finding a term $\hat{\bar{w}}_e$ that maximizes the translation probability,

$$\begin{aligned} \hat{\bar{w}}_e &= \mathrm{argmax}_{\bar{w}_e} \mathrm{Pr}(\bar{w}_e|\bar{w}_f) \\ &= \mathrm{argmax}_{\bar{w}_e} \sum_{m=1}^{M} \lambda_m h_m(\bar{w}_e, \bar{w}_f), \quad (2) \end{aligned}$$

where we have M feature functions $h_m(\bar{w}_e, \bar{w}_f)$ and model parameters λ_m.

In addition to the typical features for the SMT framework, we introduce two features: *character-based similarity*, and *additional bilingual lexicon*. We define a *character-based similarity* feature,

$$h_{\mathrm{char\_sim}}(\bar{w}_e, \bar{w}_f) = 1 - \frac{\mathrm{ED}(\bar{w}_e, \bar{w}_f)}{\max(\bar{w}_e, \bar{w}_f)}, \quad (3)$$

where $\mathrm{ED}(x, y)$ represents a Levenshtein distance of characters between the two terms x and y[3]. We also define an *additional bilingual lexicon* feature,

$$h_{\mathrm{add\_lex}}(\bar{w}_e, \bar{w}_f) = \sum_i \log p'(\bar{w}_e^{(i)}|\bar{w}_f^{(i)}), \quad (4)$$

where $\bar{w}_e^{(i)}$ and $\bar{w}_f^{(i)}$ represent an i-th translated phrase pair on the term pair $(\bar{w}_e, \bar{w}_f)$ during the decoding, and $p'(\bar{w}_e^{(i)}|\bar{w}_f^{(i)})$ represents the phrase translation probabilities derived from the additional lexicon. The probability $p'(\bar{w}_e^{(i)}|\bar{w}_f^{(i)})$ is calculated using the maximum likelihood estimation.

3 Experiment

3.1 Data

For building a Chinese-to-Japanese lexicon, we used the Japanese-English lexicon released by JST[4] (527,206 term pairs), and the Chinese-English lexicon compiled by Wanfang Data[5] (525,259 term pairs). Both cover a wide range of named entities and technical terms that may not be included in an ordinary dictionary. As an additional lexicon, we used the Japanese-English-Chinese trilingual lexicon[6] (596,967 term pairs) generated from EDR[7] Japanese-English lexicon.

We lower-cased and tokenized all terms by the following analyzers: JUMAN[8] for Japanese, the MEMM-based POS tagger[9] for English, and *cjma* (Nakagawa and Uchimoto, 2007) for Chinese.

3.2 The sizes and coverage of merged lexicons

Table 1 shows the distinct numbers of terms in the original and merged lexicons, and the *uti-*

[3] We regard the different shapes of Han characters between Chinese and Japanese as identical in our experiments.

[4] Japan Science and Technology Agency (JST) http://pr.jst.go.jp/others/tape.html

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

[6] This data was manually compiled by NICT, Japan.

[7] http://www2.nict.go.jp/r/r312/EDR/index.html

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

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

Lexicon	L_C size	L_E size	L_J size
L_C–L_E	375,990	429,807	-
L_E–L_J	-	418,044	465,563
L_E (distinct)	-	783,414	-
Additional lex.	94,928	-	90,605
Exact matching	98,537 (26.2%)	68,996	103,437 (22.2%)
Unique matching	4,875 (1.3%)	4,875	4,875 (1.0%)

Table 1: The statistics of lexicons

lization ratio[10] in the parentheses. For comparison, we prepared two baseline systems for building Chinese-Japanese lexicons. *Exact matching* connects source and target terms that share at least one common translation term in the pivot language. *Unique matching* is an extreme approach for avoiding negative effects of polysemous pivot terms: it connects source and target terms if source, pivot, and target terms appear only once in the corresponding lexicons.

Exact matching achieved 26.2% of the utilization ratio in Japanese-to-Chinese translation, and 22.2% in Chinese-to-Japanese translation. These figures imply that about 75% of the terms remained unused in building the Japanese-Chinese lexicon. With *unique matching*, as little as 1% of Japanese and Chinese terms could be used. In contrast, our method could cover 72.8% of Chinese terms by generating Japanese terms, which was a drastic improvement in the utilization ratio.

3.3 Generating Japanese translations of the Chinese-English lexicon

For evaluating the correctness of the merged lexicons, we assumed the lexicon generated by the *unique matching* as a development/test set. Development and test sets consist of about 2,400 term pairs, respectively. Next, we input Chinese terms in the development/test set into our system based on Moses (Koehn et al., 2007), and obtained the Japanese translations. We evaluated the performance by using BLEU, NIST, and accuracy measures. Table 2 shows the evaluation results on the test set. Our system could output correct translations for 68.5% of 500 input terms. The table also reports that additional features were effective in improving the performance.

We also conducted another experiment to generate Japanese translations for Chinese terms included in an external resource. We randomly extracted 500 Chinese-English term pairs from the Wanfang Data lexicon, for which the English term cannot be mapped by the JST lexicon, but can be mapped by another lexicon *Eijiro*[11]. Table 3 shows the results for these 500 terms. *Prec1* or *Prec10* are the precisions that the 1- or 10-best translations include the correct one, respectively. *MRR* (mean reciprocal rank) is $(1/500)\sum_i(1/r_i)$, where r_i is the highest rank of the correct translations for the i-th term.

Features	BLEU	NIST	Acc.
Typical features	0.4519	7.4060	0.676
w/ *character similarity*	0.4670	7.4963	0.682
w/ *additional lexicon*	0.4800	7.5907	0.674
All	0.4952	7.7046	0.685

Table 2: Translation performance on the test set

Features/Models	Prec1	Prec10	MRR
Typical features	0.142	0.232	0.1719
w/ *character similarity*	0.136	0.224	0.1654
w/ *additional lexicon*	0.140	0.230	0.1704
All	0.140	0.230	0.1714
E-to-J translation	0.090	0.206	0.1256

Table 3: Evaluation results for the *Eijiro* dictionary

Since the input lexicons are Chinese-English term pairs, their Japanese translations can be generated directly from the English terms by applying an English-Chinese translation system. We compared our system to an English-Japanese phrase-based SMT system (*E-to-J translation*), constructed from the JST Japanese-English lexicon. Table 3 shows that our system outperformed the English-to-Japanese direct translation system.

Table 4 displays translation examples. The first example shows that our system could output a correct translation (denoted by [T]); and the E-to-J system failed to translate the source term ([F]), because it could not reorder the source English words and translate the word *pubis* correctly. In the second example, our system could reproduce Chinese characters "流体 (fluid)", but the E-to-J system output a semantically acceptable but awkward Japanese term. In the last example, the word segmentation of the source Chinese term was incorrect ("中间 腰(lumber) 淋巴(lymph) 结" is correct). Thus, our system received an invalid word "腰淋" and could not find a translation for the word.

[10]The number of terms in the original lexicon used for building the merged lexicon.

[11]http://www.eijiro.jp/

English	Chinese	Japanese (*Eijiro*)	Japanese (C-to-J)	Japanese (E-to-J)
symphysis pubis	耻骨 联合	恥骨結合	恥 骨 結合 [T]	結合 恥 (symphysis shame) [F]
ideal fluid dynamics	理想 流体 动力学	理想流体力学	理想 流体 力学 [T]	理想 液 (fluid) 力学 [F]
intermediate lumbar lymph nodes	中间 腰淋巴结	中間腰リンパ節	中 間 節 腰 淋 (intermediate node [lumbar-lymph]$_{\text{INVALID}}$) [F]	中間 腰 リンパ 節 [T]

Table 4: Translation examples on *Eijiro* dictionary

4 Conclusion

This paper proposed a novel method for building a bilingual lexicon by using a pivot language. Given two bilingual lexicons L_f–L_p and L_p–L_e, we constructed a phrase-based SMT system from L_f–L_e by merging the lexicons into a phrase translation table L_f–L_e. The experimental results demonstrated that our method improves the utilization ratio of given lexicons drastically. We also showed that the pivot approach was more effective than the SMT system that translates from L_p to L_e directly.

The future direction would be to introduce other resources such as the parallel corpora and other pivot languages into the SMT system for improving the precision and the coverage of the obtained lexicon. We are also planning on evaluating a machine translation system that integrates our model.

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).

References

Brown, Peter F., 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.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics, demonstration session*, pages 177–180.

Nakagawa, Tetsuji 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.

Nie, Jian-Yun, 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.

Och, Franz Josef and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302.

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

Schafer, Charles 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.

Tanaka, Kumiko 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.

Utiyama, Masao 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.

Wu, Hua 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.

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.