Translating Chinese Unknown Words by Automatically Acquired Templates

Ming-Hong Bai^{1,2} Yu-Ming Hsieh^{1,2} Keh-Jiann Chen¹ Jason S. Chang² 1 Institute of Information Science, Academia Sinica, Taiwan

2 Department of Computer Science, National Tsing-Hua University, Taiwan

mhbai@sinica.edu.tw, morris@iis.sinica.edu.tw,

kchen@iis.sinica.edu.tw, jason.jschang@gmail.com

Abstract

In this paper, we present a translation template model to translate Chinese unknown words. The model exploits translation templates, which are extracted automatically from a word-aligned parallel corpus, to translate unknown words. The translation templates are designed in accordance with the structure of unknown words. When an unknown word is detected during translation, the model applies translation templates to the word to get a set of matched templates, and then translates the word into a set of suggested translations. Our experiment results demonstrate that the translations suggested by the unknown word translation template model significantly improve the performance of the Moses machine translation system.

1 Introduction

Automatic translation of unknown words is still an open problem. As a result, most statistical machine translation (SMT) systems treat such words as unknown tokens and leave them untranslated. (Koehn et al., 2003; Chiang, 2005; Koehn et al., 2007)

The unknown word translation problem has generated considerable interest in recent years. Some works (e.g., Callison-Burch et al., 2006; Marton et al., 2009; Mirkin et al., 2009) focus on finding in-vocabulary paraphrases, which are then used as bridges to translate target unknown words. Li and Yarowsky (2008) proposed an unsupervised method for extracting the mappings from Chinese abbreviations and their full-forms. The method exploits the full-forms as bridges to translate the abbreviations. A prerequisite of the above methods is that the unknown words must have paraphrases (or full-forms). However, many types of unknown words do not have paraphrases (full-forms) naturally.

In contrast to paraphrasing methods, Huang et al. (2011) developed a sublexical translation method that translates an unknown word by combining the translations of its sublexicals. However, to deal with the reordering problem, model combines the translations the of sublexicals by considering both straight and inverse directions and uses a language model to select the better one. The ordering is generally morphological structure dependent, but language models only select the most fluent order without considering morphological constraints.

In this paper, we propose a translation template model to translate Chinese unknown words. Our model has a number of advantages. First, the translation templates can be extracted automatically from a word-aligned parallel corpus. Second, the word order information is encoded in the templates, so we can compose the translation of an unknown word in a more reliable order. Finally, the expansion of the nonterminal symbol in the translation templates is flexible.

The remainder of this paper is organized as follows. In the next section, we introduce the proposed translation template model. In Section 3, we describe the experimental setup; and in Section 4, we evaluate the translations of unknown words derived by our model. Section 5 contains some concluding remarks.

2 Translation Template Model

The form of a translation template is similar to that of the *hierarchical phrase pair* rule (Chiang, 2005), except that the translation template is designed for translating unknown words, whereas the hierarchical phrase pair rule is designed for translating phrases. As a result, they differ in terms of the training process and rule fitting process.

Na → < [Na ₁]業,	[Na ₁] industry>
Nc → < [Nc ₁] \overline{n} ,	$[Nc_1]$ city>
Na→< 副[Na ₁], dep	outy [Na ₁]>
Nc→<[Nv ₁] \exists , se	cretary for [Nv ₁]>

Figure 1. Examples of translation templates; the notations in brackets represent the non-terminal symbols.

As shown in Figure 1., a translation template is comprised of three parts: a non-terminal symbol (Na) on the left-hand side, a source language template ($[Na_1]$) in the middle, and a target language template ($[Na_1]$ industry) on the right-hand side.

2.1 Definition of Translation Template

Based on the symbols used by Chiang (2005), we define a translation template as follows:

$$X \to <\gamma, \alpha, \sim> \tag{1}$$

where X is a left-hand side non-terminal symbol, which is usually a part-of-speech that constrains the part-of-speech of the target unknown word; γ is a translation template of the source language, and may contain terminal and non-terminal symbols; α is a translation template of the target language, and may also contain terminal and nonterminal symbols; and \sim is a one-to-one correspondence between non-terminal occurrences in γ and non-terminal occurrences in α .

2.2 Translation Process

The steps of the translation process for a given unknown word are as follows:

- Apply translation templates to the unknown word and return the matched templates.
- Translate the word based on the matched templates.
- Compute the scores for each translation candidate.

We take "出口業" (export industry) as an example to illustrate the translation process. First, translation templates are applied to the word and a set of templates are returned (shown as Figure 2).

Na → < [Nv ₁]業	, $[Nv_1]$ industry>
Na→< [Nv ₁]業	, $[Nv_1]$ business >

```
Figure 2. The matched translation templates.
```

Then, the non-terminal symbol of each rule is expanded with the translation equivalents of the in-vocabulary word " $\boxplus \square$ " (export) and the following translation candidates are generated by the matched translation templates (shown as Figure 3).

Na → < 出口業	, export industry>
Na → < 出口業	, exportation industry>
Na → < 出口業	, export business>
Na→ < 出口業	, exportation business >
Figure 3. Translation candidates.	

In the final step, we compute each translation candidate's score, and then rank all the candidates to drive the top-n translations.

2.3 Translation Probability and Lexical Weighting

Most phrase-based SMT systems use the *translation probability* and the *lexical weighting* as the parameters of scoring functions for translated phrases (Koehn et al., 2003). The original SMT translation probability is defined as follows:

$$p(\mathbf{f} | \mathbf{e}) = \frac{freq(\mathbf{f}, \mathbf{e})}{freq(\mathbf{e})}$$
(2)

where e denotes a phrase in the source language, f denotes a phrase in the target language, and $freq(\bullet)$ denotes the frequency function.

Due to the lack of unknown words in the training data, we approximate the translation probability by using the *rule fitting probability*, which is defined as follows:

$$p(\mathbf{f} | \mathbf{e}) \cong p(X \to \langle \gamma, \alpha, \sim \rangle | \mathbf{e})$$
(3)

In our experiments, we utilized the maximum entropy model (Berger et al., 1996) to model the *rule fitting probability*. It is also difficult to estimate the *lexical weighting* for the translation candidates of an unknown word. The original lexical weighting is defined as follows:

$$p_{w}(\mathbf{f} | \mathbf{e}, a) = \prod_{i=1}^{n} \frac{1}{|\{j|(i, j) \in a\}|} \sum_{\forall (i, j) \in a} p(f_{i} | e_{j})$$
(4)

where f_i denotes a word in the source phrase, and e_i denotes the words in the target phrase.

For convenience, we assume that the alignment units of the unknown words are Chinese characters, and that the alignments

between Chinese characters and English words are fully linked. Under this assumption, the lexical weighting can be simplified as follows:

$$p_{w}(\mathbf{f} | \mathbf{e}, a) \cong \prod_{\forall f_{i} \in \mathbf{f}} \frac{1}{|\mathbf{e}|} \sum_{\forall c_{j} \in \mathbf{e}} p(f_{i} | c_{j}) \quad (5)$$

where c_i denotes a character in the source phrase **e** (a Chinese unknown word), and f_i denotes a word in the target phrase **f** (an English phrase).

2.4 Extraction of Translation Templates

The translation templates are automatically extracted from a word-aligned corpus by the following steps:

- Mark the known translation equivalents in corresponding phrase pairs in the word-aligned corpus.
- Transform the marked translation equivalent pairs into the translation template form.
- Collect the translation templates derived in the previous step and compute the frequency for each rule.

In the first step, to mine translation templates from the word-aligned corpus, we utilized multisyllabic Chinese compound words to derive translation templates by marking their translation equivalents in the word-aligned corpus, as shown in Figure 4.

POS	Chinese f	Aligned English e
Na	[旅遊]業	[tourism] industry
Na	副[廠長]	deputy [director]
Nc	[運輸]司	secretary for [transport]

Figure 4. Examples of word-aligned pairs (**f**, **e**) with marked translation equivalets in the square brackets.

In the second step, we transform the marked items into the translation template form by replacing the marked words/morphemes with non-terminal symbols. The symbols on the lefthand side are part-of-speech constraints on the unknown word. Figure 5 shows the translation templates derived from the word-aligned pairs in Figure 4.

Na	\rightarrow	< [Na ₁]業 , [Na ₁] industry >
Na	\rightarrow	$< 副[Na_1]$, deputy [Na_1]>
Nc	\rightarrow	$< [Nv_1] \exists$, secretary for $[Nv_1] >$

Figure 5. The translation templates transformed from the word-aligned pairs in Figure4.

Finally, we collect the translation templates from the translation template tagged corpus and remove low frequency templates from the list.

2.5 Rule Fitting Probability

We employ the Maximum Entropy Toolkit (Zhang, 2004) to construct the rule fitting probability model, which uses the features shown in Figure 6.

POS of the word.
First character of the word.
Last character of the word.
Each character of the word.
Length of the word.
Does the word begin with a
Chinese surname?
Does the word contain a
digital number?
POS of the previous word.
POS of the next word.
Previous word.
Next word.

Figure 6. The extra features used by the *rule fitting probability* model.

2.6 Morphological Translation Rules

Some unknown words cannot be composed with simple morphemes. For example, "百分之八十" (80 percent) has a numeric morpheme, "八十" (80), which is not enumerable. The template model is flexible to be extended to use morphological translation rules instead of translation table to generate the translations of morphemes. We use two types of morphological translation rules: *numerical* and *phonetic* morphological translation rules.

3 Experimental Setting

We evaluate the model on Moses (Koehn et al., 2007) by embedding the translations of the unknown words to test data as suggestion translations.

3.1 Baseline SMT System and Data Sets

We used the Hong Kong Parallel Text (LDC2004T08) as the training data for the Moses SMT system and our template model. The Chinese sentences were pre-processed by the CKIP Chinese word segmentation system (Ma and Chen, 2003). The language model was trained on the English Gigaword corpus (LDC2003T05). We randomly selected 340

sentences from the NIST MT08 test data as our development set, the NIST MT06 test data and the rest of the NIST MT08 as our test set.

3.2 Training

The parallel text was word-aligned by the GIZA++ toolkit (Och and Ney, 2003). Then, we utilized the word-aligned corpus to extract translation templates. This process yielded a set of translation templates and a translation template tagged corpus, which was used to train the fitting probability model. To evaluate the fitting probability model, the translation template tagged corpus was randomly split into two parts to obtain a translation template tagged training set (about 1,800,000 sentences) and a translation template tagged test set (about 200,000 sentences).

We used the translation template tagged training set to train the rule fitting probability model. Then, we used the translation template tagged test set as a pseudo gold standard to evaluate the performance of the rule fitting probability model.

We also rebuilt the experiments based on the FBIS Parallel Text (LDC2003E14), which contains about 300,000 parallel sentences to verify the stability of our model. The rebuilding process is the same as that for the Hong Kong Parallel Text.

4 Experimental Results

We evaluated the translation template model on the NIST MT06 test set and NIST 08 subset. During the evaluation, the test sets were translated by the Moses SMT system with/without the embedded translation suggestions derived by the translation template model. The parameters in Moses were tuned by minimum-error-rate training (Och, 2003) on the development set.

	MT06	MT08_sub
Baseline	23.36	19.36
Trans. table	23.47 (+0.11)	19.46 (+0.10)
Phonetic	23.83 (+0.47)	19.65 (+0.29)
Numeric	23.43 (+0.07)	19.44 (+0.08)
All	23.89 (+0.53)	19.80 (+0.44)

Table 1. Evaluation results based on the Hong Kong Parallel Text.

As mentioned in Section 2.6, the translation templates can be flexible expanded by translation table as well as by morphological translation rules. In our experiments, we exploit phonetic and numerical morphological translation rules to generate translations of morphemes. Table 1 shows the performances of the translation results with/without unknown word translation suggestions. As it shows, all of the translation expansion methods significantly improved the underlying SMT system.

To verify the stability of this method, we also rebuilt a baseline system and an unknown word translation model based on the FBIS parallel corpus, as shown in Table 2.

	MT06	MT08_sub
Baseline	24.38	19.94
Trans. table	24.54 (+0.16)	20.21 (+0.27)
Phonetic	24.78 (+0.40)	20.28 (+0.34)
Numeric	24.64 (+0.26)	20.09 (+0.15)
All	25.09 (+0.71)	20.65 (+0.71)

Table 2. Evaluation results based on the FBIS parallel corpus.

The improvement in the BLEU score is statistically significant (p < 0.01) under the paired bootstrap re-sampling test (Koehn, 2004). The experimental results show that the proposed translation template model significantly improves the performance of the statistical machine translation system.

5 Conclusion

We have proposed a method that utilizes a translation template model to translate Chinese unknown words. The translation templates can be automatically extracted from a word-aligned parallel corpus and evaluated without using extra information. Experimental results show that the model can suggest accurate unknown word translations for an existing SMT system and improve the translation quality.

References

- Berger, Adam L., Stephen A. Della Pietra, Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.
- Callison-Burch, Chris, Philipp Koehn, Miles Osborne. 2006. Improved Statistical Machine Translation Using Paraphrases. In *Proc. of HLT/NAACL 2006*. pp. 17-24
- Chiang, David. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proc. of ACL 2005.* pp. 263-270.

- Huang, Chung-chi, Ho-ching Yen and Jason S. Chang. 2011. Using Sublexical Translations to Handle the OOV Problem in Machine Translation. ACM Transactions on Asian Language Information Processing, 10(3): Article 16.
- Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In Proc. of HLT/NAACL '03. pp. 127-133.
- Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proc. EMNLP*'04. pp. 388-395.
- Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proc. of the ACL 2007 Demo and Poster.*
- Li, Zhifei and David Yarowsky. 2008. Unsupervised translation induction for Chinese abbreviations using monolingual corpora. In *Proc. of ACL 2008.* pp. 425-433.
- Ma, Wei-Yun and Keh-Jiann Chen. 2003. Introduction to CKIP Chinese word segmentation system for the first international Chinese word segmentation bakeoff. In *Proc. of the ACL Workshop on Chinese Language Processing*. pp. 168-171.
- Marton, Yuval, Chris Callison-Burch and Philip Resnik. 2009. Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases. In *Proc. of ACL/AFNLP 2009*. pp. 381-390.
- Mirkin, Shachar, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman and Idan Szpektor. 2009. Source-Language Entailment Modeling for Translating Unknown Terms. In *Proc. of ACL/AFNLP 2009*. pp. 791-799.
- Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL* 2003. pp. 160-167.
- Och, Franz Josef and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, volume 29, number 1, pp. 19-51.
- Zhang, Le. 2004. Maximum entropy modeling toolkit for python and c++. available at http://homepages.inf.ed.ac.uk/lzhang10/maxent_too lkit.html.