Chinese Unknown Word Translation by Subword Re-segmentation

Ruiqiang Zhang[1,2] and **Eiichiro Sumita**[1,2]
[1]National Institute of Information and Communications Technology
[2]ATR Spoken Language Communication Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan
{ruiqiang.zhang, eiichiro.sumita}@{nict.go.jp, atr.jp}

Abstract

We propose a general approach for translating Chinese unknown words (UNK) for SMT. This approach takes advantage of the properties of Chinese word composition rules, i.e., all Chinese words are formed by sequential characters. According to the proposed approach, the unknown word is re-split into a subword sequence followed by subword translation with a subword-based translation model. "Subword" is a unit between character and long word. We found the proposed approach significantly improved translation quality on the test data of NIST MT04 and MT05. We also found that the translation quality was further improved if we applied named entity translation to translate parts of unknown words before using the subword-based translation.

1 Introduction

The use of phrase-based translation has led to great progress in statistical machine translation (SMT). Basically, the mechanism of this approach is realized by two steps:training and decoding. In the training phase, bilingual parallel sentences are preprocessed and aligned using alignment algorithms or tools such as GIZA++ (Och and Ney, 2003). Phrase pairs are then extracted to be a phrase translation table. Probabilities of a few pre-defined features are computed and assigned to the phrase pairs. The final outcome of the training is a translation table consisting of source phrases, target phrases, and lists of probabilities of features. In the decoding phase, the translation of a test source sentence is made by reordering the target phrases corresponding to the source phrases, and searching for the best hypothesis that yields the highest scores defined by the search criterion.

However, this mechanism cannot solve unknown word translation problems. Unknown words (UNK) point to those unseen words in the training or non-existing words in the translation table. One strategy to deal with translating unknown words is to remove them from the target sentence without translation on assumption of fewer UNKs in the test data. Of course, this simple way produces a lower quality of translations if there are a lot of UNKs in the test data, especially for using a Chinese word segmenter that produces many UNKs. The translation of UNKs need to be solved by a special method.

The translation of Chinese unknown words seems more difficult than other languages because Chinese language is a non-inflected language. Unlike other languages (Yang and Kirchhoff, 2006; Nießlen and Ney, 2000; Goldwater and McClosky, 2005), Chinese UNK translation cannot use information from stem and inflection analysis. Using machine transliteration can resolve part of UNK translation (Knight and Graehl, 1997). But this approach is effective for translating phonetically related unknown words, not for other types. No unified approach for translating Chinese unknown words has been proposed.

In this paper we propose a novel statistics-based approach for unknown word translation. This approach uses the properties of Chinese word composition rules – Chinese words are composed of one or more Chinese characters. We can split longer unknown words into a sequence of smaller units: characters or subwords. We train a subword based translation model and use the model to translate the sub-

word sequence. Thus we get the translation of the UNKs. We call this approach "subword-based unknown word translation".

In what follows, section 2 reviews phrase-based SMT. section 3 describes the dictionary-based CWS, that is the main CWS in this work. Section 4 describes our named entity recognition approach. Section 5 describes the subword-based approach for UNK translation. Section 7 describes the experiments we conducted to evaluate our subword approach for translating Chinese unknown words. Section 8 describes existing methods for UNK translations for other languages than Chinese. Section 9 briefly summarizes the main points of this work.

2 Phrase-based statistical machine translation

Phrase-based SMT uses a framework of log-linear models (Och, 2003) to integrate multiple features. For Chinese to English translation, source sentence C is translated into target sentence E using a probability model:

$$P_\Lambda(E|C) = \frac{exp(\sum_{i=1}^{M} \lambda_i f_i(C,E))}{\sum_{E'} exp(\sum_{i=1}^{M} \lambda_i f_i(C,E'))} \quad \Lambda = \{\lambda_1^M,\} \tag{1}$$

where $f_i(C,E)$ is the logarithmic value of the i-th feature, and λ_i is the weight of the i-th feature. The candidate target sentence that maximizes $P(E|C)$ is the solution.

Obviously, the performance of such a model depends on the qualities of its features. We used the following features in this work.

- Target language model: an N-gram language model is used.
- Phrase translation model $p(e|f)$: gives the probability of the target phrases for each source phrase.
- Phrase inverse probability $p(f|e)$: the probability of a source phrase for a given target phrase. It is the coupled feature of the last one.
- Lexical probability $lex(e|f,a)$: the sum of the target word probabilities for the given source words and the alignment of the phrase pairs.
- Lexical inverse probability $lex(f|e,a)$: the sum of the source word probabilities for the given target words and alignment.
- Target phrase length model $\#(p)$: the number of phrases included in the translation hypothesis.
- Target word penalty model: the number of words included in the translation hypothesis.
- Distance model $\#(w)$: the number of words between the tail word of one source phrase and the head word of the next source phrase.

In general, the following steps are used to get the above features.

1. Data processing: segment Chinese words and tokenize the English.
2. Word alignment: apply two-way word alignment using GIZA++.
3. Lexical translation: calculate word lexical probabilities.
4. Phrase extraction: extract source target bilingual pairs by means of union, intersection, et. al.
5. Phrase probability calculation: calculate phrase translation probability.
6. Lexical probability: generate word lexical probabilities for phrase pairs.
7. Minimal error rate training: find a solution to the λ's in the log-linear models.

3 Dictionary-based Chinese word segmentation

For a given Chinese character sequence, $C = c_0c_1c_2\ldots c_N$, the problem of word segmentation is addressed as finding a word sequence, $W = w_{t_0}w_{t_1}w_{t_2}\ldots w_{t_M}$, where the words, $w_{t_0}, w_{t_1}, w_{t_2}, \ldots, w_{t_M}$, are pre-defined by a provided lexicon/dictionary, which satisfy

$$\begin{aligned} &w_{t_0} = c_0\ldots c_{t_0}, \quad w_{t_1} = c_{t_0+1}\ldots c_{t_1} \\ &w_{t_i} = c_{t_{i-1}+1}\ldots c_{t_i}, \quad w_{t_M} = c_{t_{M-1}+1}\ldots c_{t_M} \\ &t_i > t_{i-1}, \quad 0 \le t_i \le N, \quad 0 \le i \le M \end{aligned}$$

This word sequence is found by maximizing the function below,

$$\begin{aligned} W &= \arg\max_W P(W|C) \\ &= \arg\max_W P(w_{t_0} w_{t_1} \dots w_{t_M}) \end{aligned} \quad (2)$$

We applied Bayes' law in the above derivation. $P(w_{t_0} w_{t_1} \dots w_{t_M})$ is a language model that can be expanded by the chain rule. If trigram LMs are used, it is approximated as

$$P(w_0)P(w_1|w_0)P(w_2|w_0w_1)\cdots P(w_M|w_{M-2}w_{M-1})$$

where w_i is a shorthand for w_{t_i} .

Equation 2 indicates the process of the dictionary-based word segmentation. Our CWS is based on it. We used a beam search algorithm because we found that it can speed up the decoding. Trigram LMs were used to score all the hypotheses, of which the one with the highest LM scores is the final output.

As the name indicates, the word segmentation results by the dictionary-based CWS are dependent on the size and contents of the lexicon. We will use three lexicons in order to compare effects of lexicon size to the translations. The three lexicons denoted as Character, Subword and Hyperword are listed below. An example sentence, 黄英春住在北京市(HuangYingChun lives in Beijing City), is given to show the segmentation results of using the lexicons.

- Character: Only Chinese single characters are included in the lexicon. The sentence is split character by character. 黄/英/春/住/在/北/京/市
- Subword: A small amount of most frequent words (10,000) are added to the lexicon. Choosing the subwords are described in section 5. 黄/英/春/住/在/北京/市
- Hyperword: A big size of lexicon is used, consisting of 100,000 words. 黄/英/春/住/在/北京市

4 Named entity recognition (NER)

Named entities in the test data need to be treated separately. Otherwise, a poor translation quality was found by our experiments. We define four types of named entities: people names (nr), organization names (nt), location names (ns), and numerical expressions (nc) such as calendar, time, and money. Our NER model is built according to conditional random fields (CRF) methods (Lafferty et al., 2001), by which we convert the problem of NER into that of sequence labeling. For example, we can label the last section's example as, "黄/B_nr 英/I_nr 春/I_nr 住/O 在/O 北/B_nt 京/I_nt 市/I_nt", where "B" stands for the first character of a NE; "I", other than the first character of a NE; "O", isolated character. "nr" and "nt" are two labels of NE.

Table 1: NER accuracy

type	Recall	Precision	F-score
nr	85.32%	93.41%	89.18%
ns	87.80%	90.46%	89.11%
nt	84.50%	87.54%	85.99%
all	84.58%	90.97%	87.66%

We use the CRF++ tools to train the models for named entity recognition[1]. The performance of our NER model was shown in Table 4. We use the Peking University (PKU) named entity corpus to train the models. Part of the data was used as test data.

We stick to the results of CWS if there are ambiguities in the segmentation boundary between CWS and NER.

The NER was used only on the test data in translations. It was not used on the training data due to the consideration of data sparseness. Using NER will generate more unknown words that cannot be found a translation in the translation table. That is why we use a subword-based translation approach.

5 Subword-based translation model for UNK translation

We found there were two reasons accounting for producing untranslatable words. The first is the size of lexicon. We proposed three size of lexicons in section 3, of which the Hyperword type uses 100,000 words. Because of a huge lexical size, some of the words cannot be learned by SMT training because of limited training data. The CWS chooses only one candidate segmentation from thousands in

[1]http://chasen.org/~taku/software/CRF++/

splitting a sentence into word sequences. Therefore, the use of a candidate will block other candidates. Hence, many words in the lexicon cannot be fully trained if a large lexicon is used. The second is our NER module. The NER groups a longer sequence of characters into one entity that cannot be translated. We have analyzed this points in the last section.

Therefore, in order to translate unknown words, our approach is to split longer unknown words into smaller pieces, and then translate the smaller pieces by using Character or Subword models. Finally, we put the translations back to the Hyperword models. We call this method subword-based unknown word translation regardless of whether a Character model or Subword model is used.

As described in Section 3, Characters CWS uses only characters in the lexicon. So there is no tricks for it. But for the Subword CWS, its lexicon is a small subset of the Hyperword CWS. In fact, we use the following steps for generating the lexicon. In the beginning, we use the Hyperword CWS to segment the training data. Then, we extract a list of unique tokens and calculate their counts from the results of segmentation. Next, we sort the list as the decreasing order of the counts, and choose N most frequent words from the top of the list. We restrict the length of subwords to three. We use the N words as the lexicon for the subword CWS. N can be changed. Section 7.4 shows its effect to translations. The subword CWS uses a trigram language model to disambiguate. Refer to (Zhang et al., 2006) for details about selecting the subwords.

We applied Subword CWS to re-segment the training data. Finally, we can train a subword-based SMT translation model used for translating the unknown words. Training this subword translation model was done in the same way as for the Hyperword translation model that uses the main CWS, as described in the beginning of Section 2.

6 Named entity translation

The subword-based UNK translation approach can be applied to all the UNKs indiscriminately. However, if we know an UNK is a named entity, we can translate this UNK more accurately than using the subword-based approach. Some unknown words can be translated by named entity translation if they are correctly recognized as named entity and fit a translation pattern. For example, the same words with different named entities are translated differently in the context. The word, "九", is translated into "nine" for measures and money, "September" for calendar, and "jiu" for Chinese names.

As stated in Section 4, we use NER to recognize four types of named entities. Correspondingly, we created the translation patterns to translate each type of the named entities. These patterns include patterns for translating numerical expressions, patterns for translating Chinese and Japanese names, and patterns for translating English alphabet words. The usages are described as follows.

Numerical expressions are the largest proportion of unknown words. They include calendar-related terms (days, months, years), money terms, measures, telephone numbers, times, and addresses. These words are translated using a rule-based approach. For example, "三点十五分", is translated into "at 3:15".

Chinese and Japanese names are composed of two, three, or four characters. They are translated into English by simply replacing each character with its spelling. The Japanese name, "安倍晋三", is translated into "Shinzo Abe".

English alphabets are encoded in different Chinese characters. They are translated by replacing the Chinese characters with the corresponding English letters.

We use the above translation patterns to translate the named entities. Using translation patterns produce almost correct translation. Hence, we put the named entity translation to work before we apply the subword translation model. The subword translation model is used when the unknown words cannot be translated by named entity translation.

7 SMT experiments

7.1 Data

We used LDC Chinese/English data for training. We used two test data of NIST MT04 and NIST MT05. The statistics of the data are shown in Table 6. We used about 2.4 million parallel sentences extracted from LDC data for training. Experiments on both the MT04 and MT05 test data used the same translation models on the same training data, but the min-

Table 2: Statistics of data for MT experiments

			Chinese	English
MT	Training	Sentences	2,399,753	
		words	49,546,231	52,746,558
MT04 LDC2006E43	Test	Sentences	1,788	
		Words	49,860	
MT05 LDC2006E38	Test	Sentences	1,082	
		Words	30,816	

Table 3: Statistics of unknown words of test data using different CWS

	Hyperword+Named entities					Hyperword	Subwords	Characters
	Numerics	People	Org.	Loc.	other			
MT04	460	146	250	230	219	650	18	2
MT05	414	271	311	146	323	680	23	2

imum error rate training was different. The MT04 and MT05 test data were also used as development data for cross experiments.

We used a Chinese word segmentation tool, Achilles, for doing word segmentation. Its word segmentation accuracy was higher than the stanford word segmenter (Tseng et al., 2005) in our laboratory test (Zhang et al., 2006).

The average length of a sentence for the test data MT04 and MT05 after word segmentation is 37.5 by using the Subword CWS, and 27.9 by using the Hyperword CWS.

Table 6 shows statistics of unknown words in MT04 and MT05 using different word segmentation. Obviously, character-based and subword-based CWS generated much fewer unknown words, but sentences are over-segmented. The CWS of Hyperword generated many UNKs because of using a large size of lexicon. However, if named entity recognition was applied upon the segmented results of the Hyperword, more UNKs were produced. Take an example for MT04. There are 1,305 UNKs in which numeric expressions amount to 35.2%, people names at 11.2%, organization names at 19.2%, location names at 17.6%, and others at 16.8%. Analysis of these numbers helps to understand the distribution of unknown words.

7.2 Effect of the various CWS

As described in section 3, we used three lexicon size for the dictionary-based CWS. Therefore, we had three CWS denoted as: Character, Subword and Hyperword. We used the three CWS in turn to do word segmentation to the training data, and then built the translation models respectively. We tested the performance of each of the translation models on the test data. The results are shown on Table 4. The translations are evaluated in terms of BLEU score (Papineni et al., 2002). This experiment was just testing the effect of the three CWS. Therefore, all the UNKs of the test data were not translated, simply removed from the results.

We found the character-based CWS yielded the lowest BLEU scores, indicating the translation quality of this type is the worst. The Hyperword CWS achieved the best results. If we relate it to Table 6, we found while the Hyperword CWS produced many more UNKs than the Character and Subword CWS, its translation quality was improved instead. The fact proves the quality of translation models play a more important role than the amount of unknown word translation. Using the Hyperword CWS can generate a higher quality of translation models than the Character and Subword CWS. Therefore, we cannot use the character and subword-based CWS in Chinese SMT system due to their overall poor performance. But we found their

Table 4: Compare the translations by different CWS (BLEU scores)

	MT04	MT05
Character	0.253	0.215
Subword	0.265	0.229
Hyperword	0.280	0.236

Table 5: Effect of subword and named entity translation (BLEU)

	MT04	MT05
Baseline(Hyperword)	0.280	0.236
Baseline+Subword	0.283	0.244
Baseline+NER	0.283	0.242
Baseline+NER+Subword	0.285	0.246

usage for UNK translation.

7.3 Effect of subword translation for UNKs

The experiments in this section show the effect of using the subword translation model for UNKs. We compared the results of using subword translation with those of without using it. We also used named entity translation together with the subword translation. Thus, we could compare the effect of subword translation under conditions of with or without named entity translation. We listed four kinds of results to evaluate the performance of our approach in Table 5 where the symbols indicate:

- *Baseline:* this is the results made by the Hyperword CWS of Table 4. No subword translation for UNKs and named entity translations were used. Unknown words were simply removed from the output.

- *Baseline+Subword:* the results were made under the same conditions as the first except all of the UNKs were extracted, re-segmented by the subword CWS and translated by the subword translation models. However, the named entity translation was not used.

- *Baseline+NER:* this experiment did not use subword-based translation for UNKs. But we used named entity translation. Part of UNKs was labeled with named entities and translated by pattern match of section 6.

- *Baseline+NER+Subword:* this experiment used the named entity translation and the subword-based translation. The difference from the second one is that some UNKs were translated by the translation patterns of section 6 at first and the remaining UNKs were translated using the subword model (the second one translated all of the UNKs using the subword model).

The results of our experiments are shown in Table 5. We found the subword models improved translations in all of the experiments. Using the subword models on the MT04 test data improved translations in terms of BLEU scores from 0.280 to 0.283, and from 0.236 to 0.244 on the MT05 test data. While only small gains of BLEU were achieved by UNK translation, this improvement is sufficient to prove the effectiveness of the subword models, given that the test data had only a low proportion of UNKs.

The BLEU scores of "Baseline+NER" is higher than that of "Baseline", that proves using named entity translation improved translations, but the effect of using named entity translation was worse than using the subword-based translation. This is because the named entity translation is applicable for the named entities only. However, the subword-based translation is used for all the UNKs.

When we applied named entity translation to translate some of recognized named entities followed by using the subword models, we found BLEU gains over using the subword models uniquely, 0.2% for MT04 and 0.2% for MT05. This experiment proves that the best way of using the subword models is to separate the UNKs that can be translated by named entity translation from those that cannot, and let the subword models handle translations of those not translated.

Analysis using the bootstrap tool created by Zhang et al. (Zhang et al., 2004) showed that the results made by the subword translations were significantly better than the ones not using it.

7.4 Effect of changing the size of subword lexicon

We have found a significant improvement by using the subword models. The essence of the approach

Table 6: BLEU scores for changing the subword lexicon size

subword size	MT04	MT05
character	0.280	0.237
10K	0.283	0.244
20K	0.283	0.240

is to split unknown words into subword sequences and use subword models to translate the subword sequences. The choices are flexible in choosing the number of subwords in the subword lexicon. If a different subword list is used, the results of the subword re-segmentation will be changed. Will choosing a different subword list have a large impact on the translation of UNKs? As shown in Table 6, we used three classes of subword lists: character, 10K subwords and 20K subwords. The "character" class used only single-character words, about 5,000 characters. The other two classes, "10K" and "20K", used 10,000 and 20,000 subwords. The method for choosing the subwords was described in Section 5. We have used "10K" in the previous experiments. We did not use named entity translation for this experiment.

We found that using "character" as the subword unit brought in nearly no improvement over the baseline results. Using 20K subwords yielded better results than the baseline but smaller gains than that of using the 10K subwords for MT05 data. It proves that using subword translation is an effective approach but choosing a right size of subword lexicon is important. We cannot propose a better method for finding the size. We can do more experiments repeatedly to find this value. We found the size of 10,000 subwords achieved the best results for our experiments.

8 Related work

Unknown word translation is an important problem for SMT. As we showed in the experiments, appropriate handling of this problem results in a significant improvement of translation quality. As we have known, there exists some methods for solving this problem. While these approaches were not proposed in aim to unknown word translation, they can be used for UNK translations indirectly.

Most existing work focuses on named entity translation (Carpuat et al., 2006) because named entities are the large proportion of unknown words. We also used similar methods for translating named entities in this work.

Some used stem and morphological analysis for UNKs such as (Goldwater and McClosky, 2005). Morphological analysis is effective for inflective languages but not for Chinese. Using unknown word modeling such as backoff models was proposed by (Yang and Kirchhoff, 2006).

Other proposed methods include paraphrasing (Callison-Burch et al., 2006) and transliteration (Knight and Graehl, 1997) that uses the feature of phonetic similarity. However, This approach does not work if no phonetic relationship is found.

Splitting compound words into translatable subwords as we did in this work have been used by (Nießlen and Ney, 2000) and (Koehn and Knight, 2003) for languages other than Chinese where detailed splitting methods are proposed. We used forward maximum match method to split unknown words. This splitting method is relatively simple but works well for Chinese. The splitting for Chinese is not as complicated as those languages with alphabet.

9 Discussion and conclusion

We made use of the specific property of Chinese language and proposed a subword re-segmentation to solve the translation of unknown words. Our approach was tested under various conditions such as using named entity translation and varied subword lexicons. We found this approach was very effective. We are hopeful that this approach can be applied into languages that have similar features as Chinese, for example, Japanese.

While the work was done on a SMT system which is not the state-of-the-art [2], the idea of using subword-based translation for UNKs is applicable to any systems because the problem of UNK translation has to be faced by any system.

Acknowledgement

The authors would like to thank Dr.Michael Paul for his assistance in this work, especially for evaluating methods and statistical significance test.

[2]The BLEU score of the top one system is about 0.35 for MT05 (http://www.nist.gov/speech/tests/mt/).

References

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *HLT-NAACL-2006*.

Marine Carpuat, Yihai Shen, Xiaofeng Yu, and Dekai Wu. 2006. Toward Integrating Word Sense and Entity Disambiguation into Statistical Machine Translation. In *Proc. of the IWSLT*.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of the HLT/EMNLP*.

Kevin Knight and Jonathan Graehl. 1997. Machine transliteration. In *Proc. of the ACL*.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *EACL-2003*.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML-2001*, pages 591–598.

Sonja Nießlen and Hermann Ney. 2000. Improving smt quality with morpho-syntactic analysis. In *Proc. of COLING*.

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

F. J. Och. 2003. Minimum error rate training for statistical machine trainslation. In *Proc. ACL*.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th ACL*, pages 311–318, Philadelphia, USA.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for Sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju, Korea.

Mei Yang and Katrin Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *EACL-2006*.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *Proceedings of the LREC*.

Ruiqiang Zhang, Genichiro Kikui, and Eiichiro Sumita. 2006. Subword-based tagging by conditional random fields for chinese word segmentation. In *Proceedings of the HLT-NAACL*.