Handling of out-of-vocabulary words in phrase-based statistical machine translation for Hindi-Japanese

Michael Paul and Eiichiro Sumita
NICT/ATR
Hikaridai 2-2-2, Keihanna Science City,
619-0288 Kyoto, Japan
Michael.Paul@nict.go.jp

Karunesh Arora
CDAC
Anusandhan Bhawan, C 56/1 Sector 62,
201-307 Noida, India
karunesharora@cdacnoida.in

Abstract

This paper proposes a method for handling out-of-vocabulary (OOV) words that cannot be translated using conventional phrase-based statistical machine translation (SMT) systems. For a given OOV word, lexical approximation techniques are utilized to identify and substitute spelling and inflectional word variants that occur in the training data. In order to increase the coverage of such word variant translations, the SMT translation model is extended by adding new phrase translations for all source language words that do not have a single-word entry in the original phrase-table, but only appear in the context of larger phrases. The effectiveness of the proposed method is investigated for the translation of Hindi-to-Japanese. The methodology is generic and can also be extended for other language pairs.

1. Introduction

Phrase-based SMT systems train their statistical models using parallel corpora. However, words that do not appear in the training corpus cannot be translated. Dealing with languages of rich morphology like *Hindi* and having a limited amount of bilingual resources makes this problem even more severe. Due to a large number of inflectional variations, many inflected words may not occur in the training corpus. For unknown words, no translation entry is available in the statistical translation model (*phrase-table*). Henceforward, these OOV words cannot be translated.

In this paper, we focus on the following two types of OOV words: (1) *words which have not appeared in the training corpus*, but for which other inflectional forms related to the given OOV can be found in the corpus, and (2) *words which appeared in the phrase-table in the context of larger phrases*, but do not have an individually phrase-table entry.

For a given OOV word, lexical approximation techniques are utilized to identify spelling and inflectional word variants that occur in the training corpus. The lexical approximation method applies spelling normalizers and lemmatizers to obtain word stems and generates all possible inflected word forms, whereby the variant candidates are chosen from the closest category sets to ensure grammatical features similar to the context of the OOV word. A vocabulary filter is then applied to the list of potential variant candidates to select the most frequent variant word form. All OOV words in the source sentence are replaced with appropriate word variants that can be found in the training corpus, thus reducing the amount of OOV words in the input (cf. Section 2.1).

However, a source word can only be translated in phrase-based SMT approaches, if a corresponding target phrase is assigned in the phrase-table. In order to increase the coverage of the SMT decoder, we extend the phrase-table by adding new phrase-pairs for all source language words that do not have a single-word entry in the phrase-table, but only appear in the context of larger phrases. For each of these source language words *SW*, a list of target words that occur in phrases aligned to source phrases containing *SW* in the original phrase-table is extracted and the longest sub-phrase of these target phrase entries is used to add a new phrase-table entry for *SW*. The extended phrase-table is than rescored to adjust the translation probabilities of all phrase-table entries accordingly (cf. Section 2.2).

The effectiveness of the proposed method is investigated for the translation of Hindi-to-Japanese. Section 3 summarizes the experimental results that are discussed in Section 4.

2. Handling of OOV Words

The proposed method addresses two independent, but related problems of OOV word translation approaches (cf. Figure 1). In the first step, each input sentence word that does not appear in the training corpus is replaced with the variant word form most frequently occuring in the training corpus, that can be generated by spelling normalization and feature inflection (cf. Section 2.1). In the second step, the phrase-table is extended by adding new phrase translation pairs for all source language words that do not have a single-word entry in the phrase-table, but only appear in the context of larger phrases (cf. Section 2.2).

2.1. Lexical Approximation

A phenomenon common to languages with rich morphology is the large number of inflectional variant word forms that can be generated for a given word lemma. In this paper, we deal with this problem by normalizing spelling variations and identifying inflectional word variations in order to reduce the number of OOV words in a given input sentence. The structure of the proposed lexical approximation method is sum-

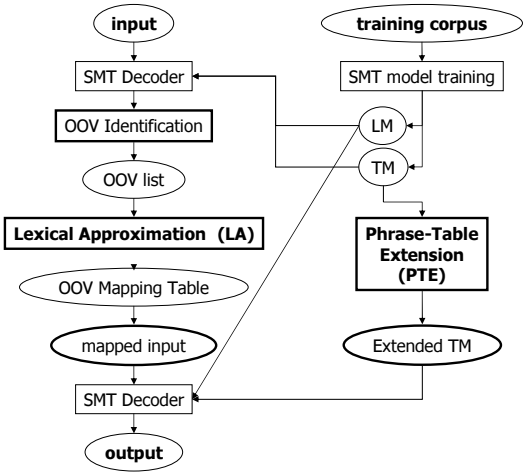

Figure 1: Outline of OOV Translation Method

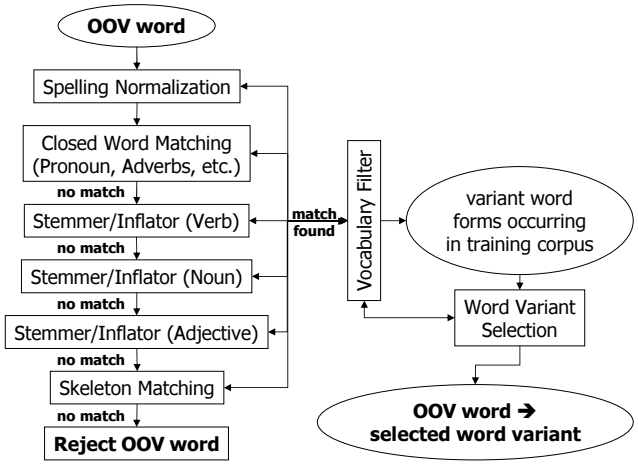

Figure 2: Lexical Approximation Method

marized in Figure 2. First, a *spelling normalizer* is applied to the input in order to map given input words to standardized spelling variants (cf. Section 2.1.1). Next, a closed word list is applied to normalize *pronouns*, *adverbs*, etc. (cf. Section 2.1.2). Content words are approximated by combining word stemming and inflectional feature generation steps for *verbs*, *nouns*, and *adjectives*, respectively (cf. Section 2.1.3). Finally, a *skeleton match* is applied (cf. Section 2.1.4). In order to identify a OOV word variant that can be translated reliably, a vocabulary filter is applied to the set of generated variant word forms, which selects the variant most frequently occuring in the training corpus.

2.1.1. Spelling Normalization

In Hindi and other Indian languages, words can be written in more than one way. Many of the spelling variations are *acceptable* variant forms. However, the lack of consistent usage of standardized writing rules resulted in *non-standard spelling variations* that are frequently used for writing. The spelling normalization module maps different word forms to one standard single word form.

2.1.2. Closed Word Matching

Words belonging to categories like *pronoun*, *adverbs*, or *postpositions appearing after nouns* belong to a closed set. These are grouped together according to grammatical feature similarities to ensure contextual meaning similarity. For examples, pronoun word forms are grouped in different categories according to their *case* or *person* attributes. The closed word form matching is applied for each category separately.

2.1.3. Stemming and Inflation

Concerning content words, two separate strategies are applied to identify variant word forms. In the first step, an OOV word is treated as an "inflected word form" and a *word stemmer* is applied to generate the corresponding root word form. In the second step, all inflectional word forms are generated from the root word according to the inflectional attributes of the respective word class. The module generates word variants for *verbs*, *nouns*, and *adjectives* separately.

2.1.4. Skeleton Matching

The final module to identify variant word forms generates the "skeletonized word form" of an OOV word by deleting dependent vowels that follow consonants. The obtained skeleton is then matched with the skeletonized word forms of the training corpus vocabulary and tyhe matched vocabulary words are treated as the OOV word variant. However, skeleton matching might result in the selection of a contextually different word, especially for OOV words of shorter length. Therefore, skeleton matching is applied, only if the other modules fail to generate any known word variant.

2.2. Phrase-Table Extension

The statistical translation model[1] of phrase-based SMT approaches consists of a source language and target language phrase pair together with a set of model probabilities and weights, that describe how likely these phrases are translations of each others in the context of sentence pairs seen in the training corpus. During decoding, the most likely phrase translation combination is selected for the translation of the input sentence [1]. Source words can only be translated in phrase-based SMT approaches, if a corresponding target phrase is assigned in the phrase-table. In order to increase the coverage of the SMT decoder, we extend the phrase-table by adding new phrase-pairs for all source language words SW that do not have a single-word entry in the phrase-table, but only appear in the context of larger phrases. The phrase-table extension method is illustrated in Figure 3.

For each of the source language words *SW* that does not have a single-word entry, all source phrases containing *SW* together with the aligned target phrases are extracted from the original phrase-table. Given these phrases, a vocabulary list T of target words sorted for occurence counts is generated. For each source word other than SW in the obtained source vocabulary list, a similar target vocabulary list is extracted and used to filter-out target word candidates in T that

[1]For details on phrase-table generation, see http://www.statmt.org/moses/?n=Moses.Background

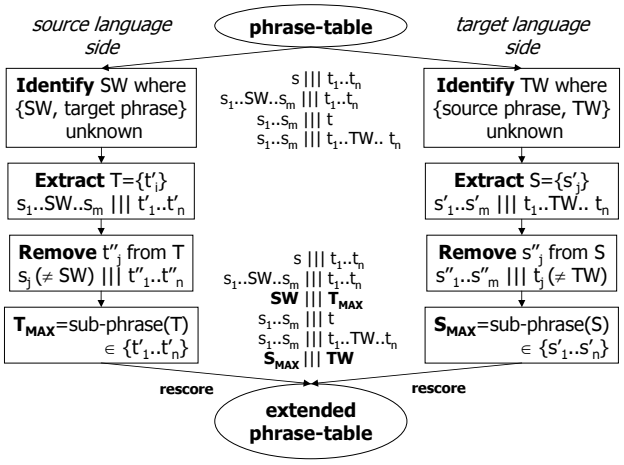

Figure 3: Phrase-Table Extension Method

cannot be aligned to SW. The remaining bag of words is than utilized to select the longest target language sub-phrase T_{MAX} of the respective original phrase-table entries and to add a new phrase-table entry $\{SW, T_{MAX}\}$. Similarily, source language translations S_{MAX} for target language words *TW* that does not have a single-word entry in the original phrase-table are obtained. The extended phrase-table is than rescored to adjust the translation probabilities of all phrase-table entries accordingly.

3. Experiments

The effectiveness of the proposed method is investigated for the translation of Hindi-to-Japanese using the *Basic Travel Expressions Corpus* (BTEC), which consists of translations of utterances in the travel domain [2]. The characteristics of the utilized BTEC corpus are summarized in Table 1.

Table 1: BTEC corpus

BTEC Corpus		Training	Evaluation
# of sentence pairs		19,972	510
Hindi	words	194,173	5,105
	vocabulary	13,681	995
	avg. length (words/sen)	9.7	8.4
Japanese	words	206,893	4,288
	vocabulary	8,609	930
	avg. length (words/sen)	10.3	8.4

For translation, an in-house phrase-based SMT decoder comparable to the open-source MOSES decoder [1] was used. For evaluation, the standard evaluation metrics BLEU[2] and METEOR[3] are applied. In addition, subjective evaluation using the *paired comparison* metrics was conducted. The output of two MT systems were given to humans who had to assign one of the following four ranks: "$better$" (first system is better than second one), "$same$" (identical MT output), "$equiv$" (different MT output, but no translation quality difference), and "$worse$" (first system is worse than second one). The *gain* of the first MT system towards the second one is calculated as the difference of the percentages of improved and degraded translations (*%better - %worse*).

[2]The *geometric mean of n-gram precision* by the system output with respect to reference texts [3].

[3]The unigram overlap between a translation and reference texts using various levels of word/stem matches [4].

3.1. Effects of Lexical Approximation

In order to investigate the effects of the proposed lexical approximation method, a standard phrase-based SMT decoder was applied to (1) the original evaluation corpus (**baseline**) and (2) the modified evaluation corpus after lexical approximation (**LA**). Table 2 shows a large reduction in OOV words of 22.8% when lexical approximation is applied. The number of input sentences containing OOV words decreased by 14.6%. Consequently, the amount of translated words increased, whereby the average sentence length of the obtained translations for sentences with OOV words increased from 8.9 to 9.6 words per sentence.

Table 2: OOV Word Reduction

	sentences with OOV	OOV words
baseline	59.2%	10.8% (442)
LA	50.0%	6.9% (341)

Table 3 summarizes the results of the automatic evaluation, whereby slightly worse BLEU scores, but improved METEOR scores were achieved for the lexical approximation method.

Table 3: Automatic Evaluation Scores for LA

	BLEU	METEOR
baseline	0.3985	0.6053
LA	0.3917	0.6105

3.2. Effects of Phrase-Table Extension

The phrase-table generated from the Hindi-Japanese training corpus contained 73,790 translation phrase pairs, whereby 5,376 source vocabulary words didn't have a single-word-entry. After the phrase-table extension, the size of the translation model increased by 7.3%.

The effects of the phrase-table extension are shown in Table 4, whereby the only difference between the systems is the usage of the original phrase-table (*baseline*) versus the extended phrase-table (*PTE*). Similarily to the lexical approximation, the BLEU scores are slightly worse, but a moderate gain is obtained for the METEOR metrics.

Table 4: Automatic Evaluation Scores for PTE

	BLEU	METEOR
baseline	0.3985	0.6053
PTE	0.3931	0.6076

3.3. Combination of LA and PTE

In order to combine both methods, we applied the lexical approximation to replace OOV words with appropriate variant

word forms in the evaluation corpus and used the extended phrase-table during SMT decoding. The automatic scores of the MT outputs are summarized in Table 5. The results show that the tendency of lower BLEU scores in contrast to higher METEOR scores still remains.

Table 5: Automatic Evaluation Scores for LA+PTE

	BLEU	METEOR
baseline	0.3985	0.6053
LA+PTE	0.3833	0.6110

In order to get an idea on how much the translation quality of a single sentence is effected by the proposed method, a subjective evaluation using *paired comparison* is applied, whereby the *baseline* system is compared to the combination of lexical approximation and phrase-table extension.

Table 6: Subjective Evaluation (Paired Comparison)

baseline vs.	TOTAL	GAIN	better	same	equiv	worse
LA+PTE	111	+ 7.2%	28.8%	17.2%	32.4%	21.6%

The results summarized in Table 6 show a large gain in translation quality. A total of 21.8% of the evaluation input sentences were addressed improving 7.2% of the translations.

Table 7 gives some examples of the subjective evaluation results. In the *better* example, the proper noun "jApAna" can be recovered successfully, thus adding important information to the translation output. In the *equivalent* example, the OOV word is wrongly translated as the sentence verb, but it does not effect the quality of the translation output, as the verb phrase was omitted in the original translation. However, in the *worse* example, the skeleton match selects a contextual different OOV word variant ("*capital*" instead of "*adult*") that changes the meaning of the translation output, thus resulting in a less acceptable translation.

4. Discussion

The experimental results in Section 3 showed that the lexical approximation and phrase-table extension methods successfully can be applied to handle OOV words, if variant word forms and appropriate phrase translation pairs are extracted from the training corpus. Conventional automatic evaluation metric scores are affected quite differently by the proposed method. The reason is that the OOV word replacement results in an increased number of translatable words. However, due to contextual shifts caused by lexical approximation and the automatic phrase-table extension, inappropriate phrase translations might be utilized to generate the final output. As the BLEU metric is quite sensitive to the word order of the translation output, scores might decrease. On the other hand, the METEOR metrics focuses more on the information expressed in the translation. Therefore, recovering unknown content words like verbs or nouns will result in higher METEOR scores, which is also reflected in the subjective evaluation results.

Table 7: Translation Examples

	[better]
input:	maiM jApAna kalekTa phona karanA chAhatA hUM .
	(*I'd like to make a collect call to Japan.*)
reference:	コレクトコールで日本にかけたいのですが。
(OOV)	"jApAna"→ [PTE] "日本" (*Japan*)
baseline:	コレクトコールをかけたいのですが。
proposed:	日本へのコレクトコールをお願いしたいのですが。
	[equivalent]
input:	kala subaha sAtha baje maiM kamarA Cho.DUMgA .
	(*I'll be checking out at seven a.m. tomorrow.*)
reference:	明日七時にチェックアウトします。
(OOV)	"subaha" → [PTE] "空く" (*to open*)
	[correct] "朝" (*morning*)
baseline	明日七時に部屋を。
proposed:	明日七時に部屋を開く。
	[worse]
input:	kRRipayA , do praudhon ke lie .
	(*Two adults , please .*)
reference:	大人二名で御願いします。
(OOV)	"praudhon" → [PTE] "首都" (*capital*)
	[correct] "大人" (*adult*)
baseline	二つを御願いします。
proposed:	首都を二つ御願いします。

5. Conclusion

In this paper, we proposed a robust and generic method to translate words not found in the training corpus by using lexical approximation and automatic phrase-table extension techniques. The experimental results for Hindi-to-Japanese revealed that the combination of both methods improved the translation quality of 7% of the input sentences containing OOV words. Further improvements can be expected when advanced phrase alignment techniques as well as external dictionaries are incorporated in order to improve the quality of additional phrase-table entries.

6. Acknowledgement

We would like to thank G. Varkey, V. N. Shukla, and S. S. Agrawal of CDAC Noida for constant support and conducive environment for this work. Special thanks are due to Prof. K. K. Goswami for providing linguistic support.

7. References

[1] P. K. et al., "Moses: Open Source Toolkit for SMT," in *Proc. of the 45th ACL, Demonstration Session*, Prague, Czech Republic, 2007, pp. 177–180.

[2] G. K. et al., "Comparative study on corpora for speech translation," *IEEE Transactions on Audio, Speech and Language*, vol. 14(5), pp. 1674–1682, 2006.

[3] K. P. et al., "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proc. of the 40th ACL*, Philadelphia, USA, 2002, pp. 311–318.

[4] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation," in *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures*, Ann Arbor, US, 2005, pp. 65–72.