# TRANSLATION OF UNKNOWN WORDS IN PHRASE-BASED STATISTICAL MACHINE TRANSLATION FOR LANGUAGES OF RICH MORPHOLOGY

*Karunesh Arora*

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

*Michael Paul and Eiichiro Sumita*

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

**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 spelling and inflectional word variants that occur in the training data. All OOV words in the source sentence are replaced with appropriate word variants that are found in the training corpus, thus reducing the amount of OOV words in the input. Moreover, in order to increase the coverage of such word 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 translations of Hindi-to-Japanese. The methodology can easily be extended for other language pairs of rich morphology.

***Index Terms***— statistical MT, out-of-vocabulary words, lexical approximation, phrase-table extension

## 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 make 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 individual phrase-table entry.

There have been some efforts in dealing with these types of OOV words. In [1], external bilingual dictionaries are used to obtain target language words for unknown proper nouns. Their training corpus is annotated for word categories like *place name*, *person name*, etc. and for each category a high-frequency word is used to (a) replace the OOV word in the input, (b) translate the modified sentence and (c) re-substitute the target language expression according to the external dictionary entries. However, this approach does not take into account any inflectional word variant context of the original OOV words. In addition, the approach depends on the coverage of the utilized external dictionary and is limited to the pre-defined categories.

In [2], orthographic features are utilized to identify lexical approximations for OOV words, but these words may be contextually different, thus resulting in wrong translations. Moreover, word translations with translation probabilities above a heuristic threshold are extracted from the Viterbi alignment of the training corpus and added to the phrase-table. However, words with alignment scores below that threshold cannot be translated.

In contrast to these previous approaches, this paper proposes a method of handling OOV words that obtains (1) finer lexical approximations due to the handling of word variations and the context of inflectional features and (2) avoids translation errors due to misaligned word pairs by exploiting phrase translations of the original phrase-table directly.

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.

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 re-scored to adjust the translation probabilities of all phrase-table entries accordingly.

The effectiveness of the proposed method is investigated for translations of Hindi-to-Japanese. However, the methodology can easily be extended for other language pairs of rich morphology.

The paper is structured as follows: Section 2 introduces the morphological features of the Hindi language. The proposed method for handling OOV words is described in detail in Section 3. Experiment results are summarized in Section 4 and are discussed in Section 5.

## 2. HINDI MORPHOLOGY

The languages of India belong to four major families: *Indo-Aryan* (a branch of the Indo-European family), *Dravidian*, *Austroasiatic* (Austric), and *Sino-Tibetan*, with the overwhelming majority of the population speaking languages belonging to the first two families. The four major families are different in their form and construction, but they share many orthographic similarities, because their scripts originate from Brahmi [3].

The *Hindi* language belongs to the Indo-Aryan language family. Hindi is spoken in vast areas of northern India and is written in Devanagari script [4]. However, two popular transliteration schemes (*ITRANS* [5] and *WX* [6]) are used for coding[1]. In Hindi, words belonging to various grammatical categories appear in lemma and inflectional forms. The inflectional forms get generated by truncating characters appearing at the end of words and adding suffixes to them, e.g., in case of *nouns*, the words are inflected based on the *number* (singular or plural), *case* (direct or oblique), and *gender* (masculine and feminine) which results in different inflectional word forms.

## 3. 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

[1]In this paper, all examples are given using the ITRANS coding scheme.

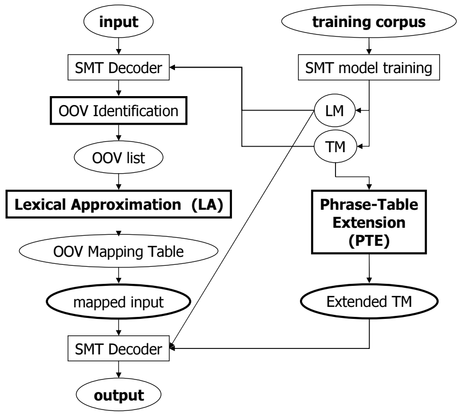


**Fig. 1**. Outline of OOV Translation Method

can be generated by spelling normalization and feature inflection (cf. Section 3.1). However, a source word can only be translated in phrase-based SMT approaches if a corresponding target phrase is assigned in the phrase-table. Therefore, 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 3.2).

### 3.1. Lexical Approximation (LA)

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 addition, allowing the flexibility of having spelling variations increases the number of correct, but different word forms in such a language. This phenomenon causes severe problems when languages of rich morphology are used as the input of a translation system, especially for languages having only a limited amount of resources available.

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 summarized 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 3.1.1). Next, a closed word list is applied to normalize *pronouns*, *adverbs*, etc. (cf. Section 3.1.2). Content words are approximated by combining word stemming and inflectional feature generation steps for *verbs*, *nouns*, and *adjectives*, respectively (cf. Section 3.1.3). Only if none of the generated variant word forms occured in the training corpus, a skeleton match is applied. Dependent vowels following consonants are removed from the OOV word and the obtained skeleton is matched against the list of all known vocabulary skeletons

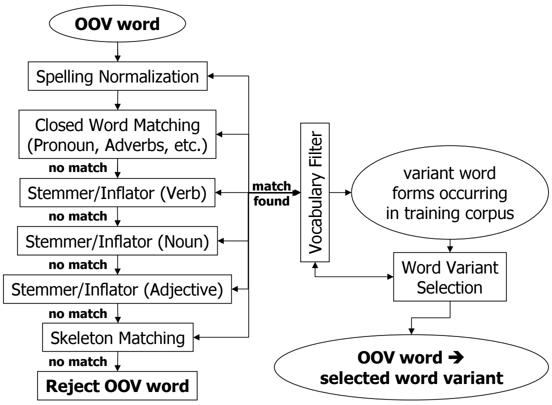


**Fig. 2**. Lexical Approximation Method

and the corresponding vocabulary is treated as a variant word form (cf. Section 3.1.4).

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.

#### *3.1.1. 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. For example, words having nasal consonants without inherent vowel sound (so-called *half-nasal consonants*) are mapped to the symbol "Anuswar" (a diacritic mark used for nasalization of consonants), e.g., "afka" ("*number*") is mapped to "aMka".

#### *3.1.2. Closed Word Matching*

Words belonging to categories like *pronoun*, *adverbs*, or *post-positions 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, e.g., the *Genitive case* variant word forms of the first-person pronoun "merA" (*my*) is "merI" in the feminine case and "mere" in the plural form. The closed word form matching is applied for each category separately. The list of all word forms passing the vocabulary filter is returned by this module.

#### *3.1.3. Stemming and Inflection*

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. Examples for the generation of inflectional forms of verbs and nouns are given in Table 1 and Table 2, respectively.

**Table 1**. Verb Inflections

| Category | "jA" (*to go*) |
|---|---|
| Present | jAtA, jAtI, jAte |
| Past | gayA, gayI, gaI, gaye, gae, gayIM |
| Future | jAU.NgA, jAegA, jAoge, jAe.Nge, jAU.NgI, jAegI, jAe.NgI |
| Subjunctive | jAU.N, jAe, jAe.N, jAo |

**Table 2**. Noun Inflections

| Case/Number | "lad.DakA" (*boy*) | "lad.DakI" (*girl*) |
|---|---|---|
| Direct/Singular | lad.DakA | lad.DakI |
| Direct/Plural | lad.Dake | lad.DakiyAz |
| Oblique/Singular | lad.Dake | lad.DakI |
| Oblique/Plural | lad.DakoM | lad.DakiyoM |

Concerning Hindi adjectives, two categories are distinguished. The *red adjectives* do not vary in form, whereas the *black adjectives* vary according to the *gender*, *number* and *case* features of the noun they precede (cf. Table 3).

**Table 3**. Adjective Inflections

| Case/Number | "kAlA"(*black*) | |
|---|---|---|
| Direct/Singular | kAlA | kAlI |
| Direct/Plural | kAle | kAlI |
| Oblique/Singular | kAle | kAlI |
| Oblique/Plural | kAle | kAlI |

#### *3.1.4. 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, e.g., the skeleton of the Hindi word "batAyA" (*told*) is "bty". The obtained skeleton is then matched with the skeletonized word forms of the training corpus vocabulary. In case of a skeleton match, the respective vocabulary word is 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, the skeleton matching module is applied only if the other modules fail to generate any known word variant.

### 3.2. Phrase-Table Extension (PTE)

The statistical translation model[2] of phrase-based SMT approaches consists of a source language and target language

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

phrase pair together with a set of model probabilities and weights, that describe how likely these phrases are translations of each other 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 [7]. 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.

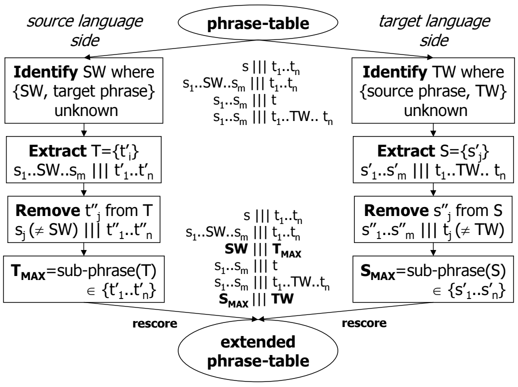


**Fig. 3**. Phrase-Table Extension Method

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 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 do not have a single-word entry in the original phrase-table are obtained. The extended phrase-table is than re-scored to adjust the translation probabilities of all entries accordingly.

## 4. EXPERIMENTS

The effectiveness of the proposed method is investigated for translations of Hindi-to-Japanese using the *Basic Travel Expressions Corpus* (BTEC), which is a collection of sentences that bilingual travel experts consider useful for people going to or coming from another country and cover utterances for potential subjects in travel situations [8]. The characteristics of the utilized BTEC corpus are summarized in Table 4.

**Table 4**. 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 the training of the statistical models, standard word alignment (GIZA++ [9]) and language modeling (SRILM [10]) tools were used. For translation, an in-house phrase-based SMT decoder comparable to the open-source MOSES decoder [7] was used. For evaluation, the automatic evaluation metrics listed in Table 5 were applied to the translation output. Previous research on MT evaluation showed that these automatic metrics correlate best with human assessment of machine translation quality [11].

**Table 5**. Automatic Evaluation Metrics

| | |
|---|---|
| BLEU: | the *geometric mean of n-gram precision* by the system output with respect to reference translations. Scores range between 0 (worst) and 1 (best) [12] |
| TER: | *Translation Edit Rate*: a edit distance metrics that allows phrasal shifts. Scores are positive with 0 being the best possible [13] |
| METEOR: | calculates unigram overlaps between a translations and reference texts using various levels of matches (*exact*, *stem*) are taken into account. Scores range between 0 (worst) and 1 (best) [14] |
| GTM: | measures the similarity between texts by using a unigram-based F-measure. Scores range between 0 (worst) and 1 (best) [15] |

In addition, subjective evaluation using the *paired comparison* metrics was conducted. The output of two MT systems were given to a human evaluator who had to assign one of the four ranks given in Table 6. 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*).

**Table 6**. Paired Comparison Evaluation Ranks

| | |
|---|---|
| *better*: | the translation quality of the first MT system output is better than the output of the second one |
| *same*: | both MT outputs are identical |
| *equiv*: | both systems generated different MT outputs, but there is no difference in translation quality |
| *worse*: | the translation quality of the first MT system output is worse than the output of the second one |

### 4.1. 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 the following input data sets:

(1) the original evaluation corpus (**baseline**)
(2) the modified evaluation corpus after lexical approximation *without skeleton matching* ($\mathbf{LA}_w$)
(3) the modified evaluation corpus after lexical approximation *with skeleton matching* ($\mathbf{LA}_s$)

Comparing the OOV reduction rates summarized in Table 7, a large reduction in OOV words can be seen when the proposed method is applied to the original evaluation corpus, i.e., 6.8% (22.8%) for the lexical approximation without (with) skeleton matching. The number of input sentences containing OOV words decreased by 5.1% (14.6%), respectively. Consequently, the amount of translated words increased, whereby the average sentence length of the obtained translations for sentences with recovered OOV words increased from 8.9 to 9.4 (9.6) words per sentence.

**Table 7**. OOV Word Reduction

| | sentences with OOV | OOV words |
|---|---|---|
| *baseline* | 59.2% | 10.8% (442) |
| $LA_w$ | 56.5% | 8.4% (412) |
| $LA_s$ | 50.0% | 6.9% (341) |

Concerning the automatic evaluation scores, slightly worse BLEU/TER scores, but improved METEOR/GTM scores were achieved for the LA method (cf. Table 8).

**Table 8**. Automatic Evaluation Scores for LA

| | BLEU | TER | METEOR | GTM |
|---|---|---|---|---|
| *baseline* | 0.3985 | 0.4994 | 0.6053 | 0.8817 |
| $LA_w$ | 0.3949 | 0.5043 | 0.6050 | 0.8825 |
| $LA_s$ | 0.3917 | 0.5126 | 0.6105 | 0.8855 |

### 4.2. 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 9. The only difference between the systems is the usage of the original phrase-table (*baseline*) versus the extended phrase-table. Similarily to the lexical approximation, the BLEU/TER scores are slightly worse, but a moderate gain is obtained for the METEOR/GTM metrics.

**Table 9**. Automatic Evaluation Scores for PTE

| | BLEU | TER | METEOR | GTM |
|---|---|---|---|---|
| *baseline* | 0.3985 | 0.4994 | 0.6053 | 0.8817 |
| $PTE$ | 0.3931 | 0.5011 | 0.6076 | 0.8876 |

### 4.3. Combination of LA and PTE

In order to combine both methods, we applied the lexical approximation without ($LA_w$) and with ($LA_s$) skeleton matching to replace OOV words with appropriate variant word forms in the evaluation corpus and used the extended phrase-table ($PTE$) during SMT decoding. The automatic scores of the MT outputs are summarized in Table 10. The results show that the tendency of worse BLEU/TER scores in contrast to improved METEOR/GTM scores still remains.

**Table 10**. Automatic Evaluation Scores for LA+PTE

| | BLEU | TER | METEOR | GTM |
|---|---|---|---|---|
| *baseline* | 0.3985 | 0.4994 | 0.6053 | 0.8817 |
| $LA_w + PTE$ | 0.3915 | 0.5056 | 0.6078 | 0.8876 |
| $LA_s + PTE$ | 0.3833 | 0.5132 | 0.6110 | 0.8925 |

However, the automatic scoring metrics are designed to judge the translation quality of the MT output on document-level, but not on sentence-level. 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 without ($LA_w + PTE$) and with ($LA_s + PTE$) skeleton matching.

**Table 11**. Subjective Evaluation (Paired Comparison)

| $baseline$ vs. | TOTAL | GAIN | better | same | equiv | worse |
|---|---|---|---|---|---|---|
| $LA_w + PTE$ | 29 | +13.8% | 24.1% | 31.1% | 34.5% | 10.3% |
| $LA_s + PTE$ | 111 | + 7.2% | 28.8% | 17.2% | 32.4% | 21.6% |

The results summarized in Table 11 show a large gain in translation quality for both types of lexical approximation. Without skeleton matching, a total of 6% of the evaluation input sentences were addressed improving 13.8% of the translations. The usage of skeleton matching increases the coverage of the proposed method (21.8% of the input sentences were addressed), but lowers the overall gains (7.2% of improved translations).

Table 12 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.

## 5. DISCUSSION

Experiment results in Section 4 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 BLEU/TER metric scores decreased slightly,

**Table 12**. Translation Examples

| | |
|---|---|
| | **[better]** |
| input: | maiM jApAna kalekTa phona karanA chAhatA hUM . |
| | (*I'd like to make a collect call to Japan.*) |
| (OOV) | "jApAna"→ [PTE] "nihon" (*Japan*) |
| baseline: | korekutokouru o kaketai no desu ga . |
| | (*I'd like to make a collect call.*) |
| proposed: | nihon e no korekutokouru o onegai shitai no desu ga . |
| | (*I'd like to have a collect call to Japan.*) |
| | **[equivalent]** |
| input: | kala subaha sAta baje maiM kamarA Cho.DUMgA . |
| | (*I'll be checking out at seven a.m. tomorrow.*) |
| (OOV) | "subaha" → [PTE] "aku" (*to open*) |
| | [correct] "asa" (*morning*) |
| baseline | ashita shichi ji ni heya o . |
| | (*I'll do the room at seven a.m. tomorrow.*) |
| proposed: | ashita shichi ji ni heya o aku . |
| | (*I'll open the room at seven a.m. tomorrow.*) |
| | **[worse]** |
| input: | kRRipayA , do praudhon ke lie . |
| | (*Two adults, please.*) |
| (OOV) | "praudhon" → [PTE] "shuto" (*capital*) |
| | [correct] "otona" (*adult*) |
| baseline | futatsu o onegai shimasu . |
| | (*For two please.*) |
| proposed: | shuto o futatsu onegai shimasu . |
| | (*The capitol for two, please.*) |

whereby the METEOR/GTM scores improved. 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 using skeleton matching and the automatic phrase-table extension, inappropriate phrase translations might be utilized to generate the final output. In addition, the probabilities assigned to the newly added phrase-translation pairs does not necessarily reflect the correct word distribution in the training corpus. As the BLEU/TER metrics are quite sensitive to the word order of the translation output, scores might decrease. On the other hand, the METEOR/GTM metrics focus more on the information expressed in the translation. Therefore, recovering unknown content words like verbs or nouns will result in higher METEOR/GTM scores, which is also reflected in the subjective evaluation results.

## 6. CONCLUSION

In this paper, we proposed a method to translate words not found in the training corpus by using lexical approximation techniques to identify known variant word forms and adjust the input sentence accordingly. The translation coverage is increased by extending the original phrase-table with phrase translation pairs for source vocabulary words without single-word entries in the original phrase-table. Experiment results for Hindi-to-Japanese revealed that the combination of both methods improved the translation quality up to 14% for input sentences containing OOV words. Further investigations will include a detailed error analysis and the application of advanced phrase alignment techniques as well as the incorporation of external dictionaries in order to improve the quality of additional phrase-table entries.

## 7. ACKNOWLEDGEMENT

We would like to thank G. Varkey, V. N. Shukla, and S. S. Agrawal of CDAC Noida and S. Nakamura of NICT/ATR for constant support and conducive environment for this work.

## 8. REFERENCES


[1] H. Okuma et al., "Introducing Translation Dictionary into phrase-based SMT," in *Proc. of MT Summit XI*, Copenhagen, Denmark, 2007, pp. 361–368.

[2] C. Mermer et al., "The TUBITAK-UEKAE SMT System for IWSLT 2007," in *Proc. of the IWSLT*, Trento, Italy, 2007, pp. 176–179.

[3] R. Ishida, "An introduction to Indic scripts," in *Proc. of the 22nd International Unicode Conference*, 2002.

[4] Wikipedia, "Devanagari," http://en.wikipedia.org/wiki/Devanagari_script, 2007.

[5] A. Chopde, "ITRANS - Indian Language Transliteration Package," http://en.wikipedia.org/wiki/ITRANS, 2006.

[6] WX, "Roman transliteration scheme for Devanagar," http://sanskrit.inria.fr/DATA/wx.html, 2007.

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

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

[9] F. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29(1), pp. 19–51, 2003.

[10] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. of ICSLP*, Denver, 2002, pp. 901–904.

[11] M. Paul, "Overview of the IWSLT 2006 Evaluation Campaign," in *Proc. of the IWSLT*, Kyoto, Japan, 2006, pp. 1–15.

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

[13] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, "A study of translation edit rate with targeted human annotation," in *Proc. of the AMTA*, Cambridge and USA, 2006, pp. 223–231.

[14] 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.

[15] Joseph Turian, Luke Shen, and I. Melamed, "Evaluation of machine translation and its evaluation," in *Proc. of the MT Summit IX*, New Orleans, USA, 2003, pp. 386–393.