IEICE TRANS. INF. & SYST., VOL.E92–D, NO.12 DECEMBER 2009

PAPER Special Section on Natural Language Processing and its Applications

Translation of Untranslatable Words — Integration of Lexical Approximation and Phrase-Table Extension Techniques into Statistical Machine Translation

Michael PAUL[†a)], **Karunesh ARORA**[††b)], *Nonmembers*, *and* **Eiichiro SUMITA**[†c)], *Member*

SUMMARY 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 then replaced with appropriate word variants found in the training corpus, thus reducing the number 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 methods is investigated for the translation of Hindi to English, Chinese, and Japanese.

key words: *statistical machine translation, out-of-vocabulary words, lexical approximation, phrase-table extension*

1. Introduction

Phrase-based SMT systems train their statistical models using parallel corpora [1], [2]. However, words that do not appear in the training corpus cannot be translated. Dealing with languages with a 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*). As a result, these OOV words cannot be translated.

There have been several efforts in dealing with OOV words to improve translation quality. In addition to parallel text corpora, external bilingual dictionaries can be exploited to reduce the OOV problem. Adding such translation pairs to the translation model will increase the coverage of the SMT system. However, no word order information can be provided for statistical distortion models that determine the order of target language phrases, because these dictionary entries might not appear in the context of translation examples. In order to avoid the word order problems of such *dictionary-based* phrase-table extension approaches, [3] annotates their training corpus for word categories like *proper nouns* 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, these approaches depend on the coverage of the utilized external dictionary and are limited to pre-defined categories.

Data sparseness problems due to inflectional variations were previously addressed by applying word transformations using stemming or lemmatization. [4] made use of word morpheme information such as lemma/part-of-speech annotations to improve translations between Spanish and English. A tighter integration of morpho-syntactic information into the translation model was proposed by [5]. Within this *factored translation model*, words are not only a token, but a vector of factors that represent different levels of annotation. Lemma and morphological information are translated separately, and this information is combined on the output side to generate the translation. However, these approaches still suffer from the data sparseness problem, since lemmata and inflectional forms never seen in the training corpus cannot be translated.

In order to generate translations for unknown words, previous approaches focused on *transliteration* methods, where a sequence of characters is mapped from one writing system into another. For example, in order to translate names and technical terms, [6] introduced a probabilistic model that replaces Japanese *katakana** words with phonetically equivalent English words. However, transliteration approaches are limited not only to certain word categories like *proper nouns*, but also to language pairs that are orthographically/phonetically closely related.

Instead of finding an equivalent target language expression for an unknown source word, the OOV problem has also been addressed by applying *lexical approximation* techniques, where the unknown word is replaced with a closely related source language word that occurred in the training data and the modified input is translated in a standard way. For example, [7] utilized orthographic features like *string similarity* to identify lexical approximations for OOV words. However, this approach does not take into account grammatical features like part-of-speech or inflectional attributes which are necessary to translate the unknown word

Manuscript received March 20, 2009.
Manuscript revised July 17, 2009.
[†]The authors are with NICT, Kyoto-fu, 619–0289, Japan.
[††]The author is with CDAC, Noida, India.
a) E-mail: michael.paul@nict.go.jp
b) E-mail: karunesharora@cdacnoida.in
c) E-mail: eiichiro.sumita@nict.go.jp
DOI: 10.1587/transinf.E92.D.2378

*A syllabary alphabet used to write down foreign names.

in the context of the given input sentence.

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.

In contrast to the above mentioned approaches, this paper proposes a method of handling OOV words that obtains (a) finely graded lexical approximations due to the handling of word variations and the context of inflectional features and (b) larger coverage of the SMT translation model by extending the phrase-table with single word entries that only appear in the context of larger phrases of the original phrase-table.

For a given OOV word, lexical approximation techniques are utilized to identify spelling and inflectional word variants that occur in the training corpus. The proposed 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 number of OOV words in the input.

However, a source word can only be translated by 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*, all target language phrases that are assigned to source language phrases containing *SW* are extracted from the original phrase-table. For all of these target phrases, the sub-phrase corresponding to *SW* is identified and the most frequent sub phrase T_{MAX} is selected in order to extend the original phrase-table with the new lexical pair (SW, T_{MAX}). The extended phrase-table is then re-scored to adjust the translation probabilities of all phrase-table entries accordingly.

The paper is structured as follows: Section 2 introduces the characteristics of the Hindi language. Section 3 describes the proposed methods for handling OOV words using lexical approximation and phrase-table extension techniques in detail. In Sect. 4, the effectiveness of the proposed methods is primarily investigated for the translation of Hindi to English. In addition, experimental results are investigated further by considering different target languages, namely Chinese and Japanese.

2. Characteristics of Hindi

The *Hindi* language belongs to the Indo-Aryan language family. Hindi is spoken in northern India and is written in the Devanagari script. However, there exists several transliteration schemes for coding. In this paper, all examples are given using the *WX* coding scheme [8]. In Hindi, words belonging to various grammatical categories appear in lemma and inflectional forms. The inflectional forms are 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 or feminine).

3. Handling of OOV Words

The proposed method addresses two independent, but related, problems of OOV word translation approaches (cf. Fig. 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 occurring in the training corpus, which can be generated by spelling normalization and feature inflection (cf. Sect. 3.1). This approach is related to work reported in [4], [5], and [7] where stemming and lemmatization are applied to full word forms in order to either enrich SMT models with additional morphological information or to identify target language expressions similar to the unknown input words. However, their approach cannot deal with word stems and inflections that do not appear in the training corpus. In contrast, the analysis of morphological information in our approach is applied to the identification of contextually similar source language words that do occur in the training corpus. Besides stem matching, all possible inflectional word forms are generated for unknown word and matched against training corpus word forms. Therefore, although the full word form or even the respective stem form might not appear in the training corpus, unknown words can be translated in the context of the

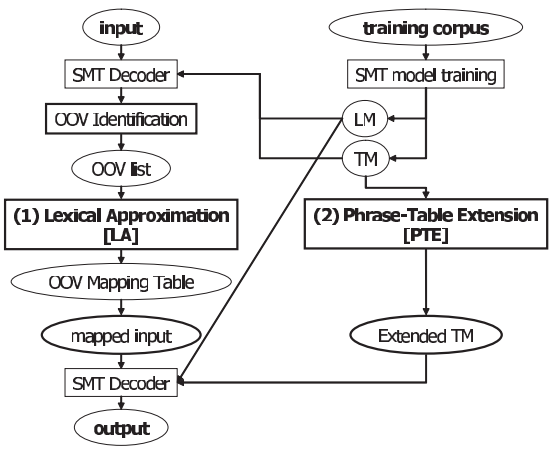

Fig. 1 Outline of the OOV translation method.

input context using related inflectional word forms occurring in the training corpus.

In addition, similar to [3], word order problems of dictionary-based approaches to OOV handling are avoided because the unknown word is translated within the context of translation examples containing the lexically approximated word form. In contrast to [3], this method is not limited to a specific word category, like *proper nouns*, but can also handle other word categories, like *common nouns* or *verbs*.

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 and only appear in the context of larger phrases (cf. Sect. 3.2). Phrase-table extension methods have been proposed previously. For example, [7] augments the phrase-table with lexical word pairs extracted from the source-to-target Viterbi alignment available from an intermediate step in the translation model training. However, due to word alignment errors, incorrect translation pairs might be extracted, which can result in the incorrect translation of an input sentence. Therefore, our approach tries to avoid translation errors due to misaligned word pairs by exploiting the context of phrase translations of the original phrase-table. In order to identify appropriate translation candidates for a source language word that does not have a single-word entry in the phrase-table, target language words that can be aligned to other source words are removed from the target phrases with the most frequent target language sub-phrase (after filtering) being selected for the translation of the source word.

3.1 Lexical Approximation (LA)

A phenomenon common to languages with a rich morphology is the large number of inflections 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 such languages are used as the input of a translation system.

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 the input. The structure of the proposed lexical approximation method is summarized in Fig. 2. In step (1), a *spelling normalizer* is applied to map input words to standardized spelling variants (cf. Sect. 3.1.1), Next, step (2) applies a closed word list to normalize *pronouns*, *adverbs*, etc. (cf. Sect. 3.1.2). In step (3), content words are approximated by combining word stemming and inflectional feature generation steps for *verbs*, *nouns*, and *adjectives*, respectively (cf. Sect. 3.1.3). Only if none of the generated variant word forms occur in the training corpus is a skeleton match applied in step (4). Dependent vowels following consonants

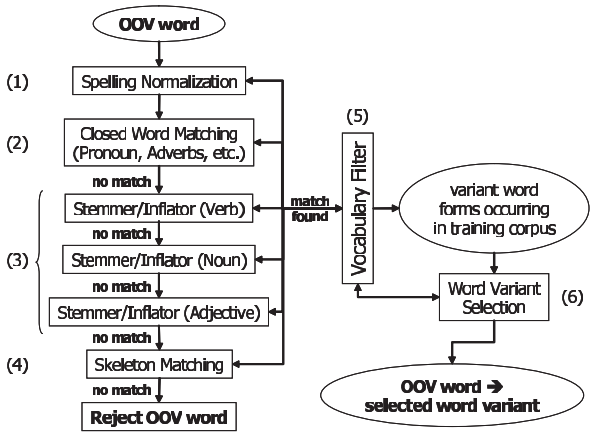

Fig. 2 Lexical approximation method.

are removed from the OOV word and the obtained skeleton is matched against the list of all known vocabulary skeletons and the corresponding vocabulary is treated as a variant word form (cf. Sect. 3.1.4). When one or more matches were found, step (5) applies a vocabulary filter in order to identify an OOV word variant by preferring candidates that belongs to the same word category as the OOV word. Finally, the variant most frequently occurring in the training corpus is selected in step (6) to replace the respective unknown word in the input sentence.

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 has resulted in *non-standard spelling variations* that are frequently used in 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 example, pronoun word forms are grouped in 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 (cf. Table 1). 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.

Table 1 Closed word matching.

Case/Number	Masculine	Feminine
Genitive/Singular	merA	merI
Genitive/Plural	mere	mere

Table 2 Stemming and inflection.

Verb	"jA" (*to go*)
Present	jAwA, jAwI, jAwe
Past	gayA, gayI, gaI, gaye, gae, gayIM
Future	jAUzgA, jAegA, jAoge, jAezge, jAUzgI, jAegI, jAezgI
Subjunctive	jAUz, jAe, jAez, jAo

Noun (Case/Num)	"ladZakA" (*boy*)	"ladZakI" (*girl*)
Direct/Singular	ladZakA	ladZakI
Direct/Plural	ladZake	ladZakiyAz

Adjective (Case/Num)	"kAlA"(*black*)	
Direct/Singular	kAlA	kAlI
Direct/Plural	kAle	kAlI

Table 3 Skeleton matching.

Full Word Form	Skeleton	Matched Word Candidates
bawAyA (*told*)	bwy	bawAye (*to tell*), ... biwAyA (*had*)
sucanA (*information*)	scn	sUcanA (*information*), socanA (*to look into*), socane (*to think*), ...

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. There are two seperate categories within Hindi adjectives. 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. Examples for the generation of inflectional forms of verbs, nouns, and adjectives are given in Table 2.

3.1.4 Skeleton Matching

The final module to identify variant word forms generates a "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. Table 3 gives some examples for word forms, its skeleton, and variant word forms that can be matched using the skeleton approach. In case of a skeleton match, the respective vocabulary word is treated as the variant of the OOV word. 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.

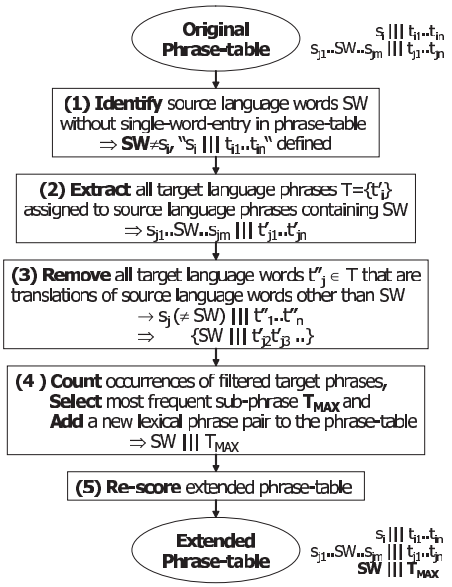

Fig. 3 Phrase-table extension method.

3.2 Phrase-Table Extension (PTE)

A standard phrase-based statistical machine translation engine uses multiple statistical models to generate a translation hypothesis in which (1) the *translation model* ensures that the source phrases and the selected target phrases are appropriate translations of each other, (2) the *language model* ensures that the target language is fluent, (3) the *distortion model* controls the reordering of the input sentence, and (4) the *word penalty* ensures that the translations do not become too long or too short. During decoding, all model scores are weighted and combined to find the most likely translation hypothesis for a given input sentence [9].

The key to a good translation is the translation model that consists of a source and target language 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 the sentence pairs seen in the training corpus [1]. However, 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 Fig. 3. In step (1), all source language words SW that do not have a single-word entry are identified in the original phrase-table. Step (2) extracts all source phrases S containing SW ($S = s_1 \dots SW \dots s_m$) together with the aligned

target phrases T ($T = t'_1 \dots t'_n$) for each SW. For each of these phrase-table pairs $\{S, T\}$, the target sub-phrase T_{SW} that corresponds to SW is obtained by removing all target words t'' from T that can be aligned to source words s_i of S other than SW according to the source-to-target Viterbi alignments available from an intermediate step in the translation model training in step (3). Based on the frequency that the obtained target language sub-phrase T_{SW} occurrs in the training corpus, the most likely target phrase T_{MAX} is selected and a new phrase-table entry $\{SW, T_{MAX}\}$ is added to the original phrase-table in step (4). Finally, step (5) rescores the extended phrase-table to adjust the probabilities of all entries accordingly.

4. Experiments

The effectiveness of the proposed method is investigated for the translation of Hindi into English, Chinese, and 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 covers utterances for potential subjects in travel situations [10]. The characteristics of the utilized BTEC corpus are summarized in Table 4. For all languages, 98,356 sentences were used for training and the evaluation data set consisted of 510 sentences.

The sentence length is given as the average number of words per sentence. The OOV word figures give the percentage of words in the evaluation data set that do not appear in the training data corpus. The corpus statistics show that languages with a rich morphology have a much larger vocabulary and a much larger number of untranslatable words in the evaluation data, i.e., the Hindi vocabulary is twice as large and the OOV rate is 10 times higher than those of the other languages. An analysis of the OOV word categories revealed that most of the OOV words in the given evaluation data sets are *common nouns* (Hindi: 42.2%, English: 53.8%, Chinese: 36.0%, Japanese: 36.0%) or *verbs* (Hindi: 21.2%, English: 25.6%, Chinese: 42.3%, Japanese: 24.9%). Only a few *proper nouns* (Hindi: 5.6%, English: 3.1%, Chinese: 2.2%, Japanese: 2.2%) are not covered by the training corpus. In order to get an idea how difficult the translation tasks may be, we calculated the language perplexity (*perpl*) of the respective evaluation data sets according to the language model used by the baseline system. The numbers in Table 4 indicate that Hindi is by far the most difficult language, followed by Chinese, English and then Japanese.

Table 4 Language resources.

Hindi	Train	Eval
vocab	24,377	1,162
avg.length	8.0	8.1
OOV	–	8.7%
perplexity	–	143.7

Chinese	Train	Eval
vocab	12,267	949
avg.length	6.7	6.8
OOV	–	0.8%
perplexity	–	26.5

English	Train	Eval
vocab	11,715	890
avg.length	7.6	7.5
OOV	–	0.8%
perplexity	–	19.9

Japanese	Train	Eval
vocab	12,317	930
avg.length	8.4	8.4
OOV	–	0.6%
perplexity	–	13.9

For the training of the SMT models, standard word alignment (GIZA++ [11]) and language model (SRILM [12]) tools were used. For the translation, an in-house phrase-based SMT decoder comparable to the open-source MOSES decoder [9] was used. For the automatic evaluation, the standard BLEU metrics [13] which calculates the *geometric mean of n-gram precision* by the system output with respect to reference translations is used. Its scores range between 0 (worst) and 1 (best).

Moreover, subjective evaluation using the *ranking* metric [14] was conducted. This evaluation method is an extension of the paired-comparison method to compare the translation quality of multiple system outputs. The task is to rank translated sentences relative to each other from "Best" to "Worst", in which ties are allowed. Three human evaluators who are native speakers of the target language were asked to grade the system outputs using a web-browser interface in which a human translation of the input was shown as the reference translation together with the MT outputs.

In order to decide whether the translation output of one MT engine is significantly better than another one, we used the *bootStrap* [15] method that (1) performs a random sampling with replacement from the *eval* data set, (2) calculates the respective evaluation metric score of each engine for the sampled test sentences and the difference between the two MT system scores, (3) repeats the sampling/scoring step iteratively[†], and (4) applies the *Student's t-test* at a significance level of 95% confidence to test whether the score differences are statistically significant. The automatic evaluation scores given in this paper are the mean scores of all *bootStrap* iterations.

Based on the above evaluation metrics, we will investigate the effectiveness of the *lexical approximation* and *phrase-table extension* methods using the Hindi-English translation outputs. Section 4.1 and Sect. 4.2 examine the effects of the LA and the PTE methods proposed in this paper as well as previous approaches (*string-similarity-based LA*, *word-alignment-based PTE*). In addition, combinations of LA and PTE techniques are evaluated in Sect. 4.3.

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 input after LA *using the string similarity matching of [7]* ($\mathbf{LA}_d$)
(3) the modified input after LA *without skeleton matching* ($\mathbf{LA}_w$)
(4) the modified input after LA *with skeleton matching* ($\mathbf{LA}_s$)

[†]We used 2000 iterations of the *bootStrap* method for the experimental results reported in this paper.

Table 5 OOV word reduction for LA.

	Sentences with OOV	OOV words
baseline	62.0% (316 sen)	11.4% (471 words)
LA_d	59.8% (305 sen)	10.7% (441 words)
LA_w	59.6% (304 sen)	10.5% (434 words)
LA_s	35.7% (182 sen)	5.2% (214 words)

Table 6 OOV word reduction for PTE.

	Sentences with OOV	OOV words
baseline	62.0% (316 sen)	11.4% (471 words)
PTE_v	54.3% (277 sen)	8.9% (367 words)
PTE_p	54.1% (276 sen)	8.8% (366 words)

Table 7 Automatic evaluation scores. (BLEU%)

baseline	36.59	LA_d+PTE_v	35.27
LA_d	37.05	LA_w+PTE_v	35.22
LA_w	37.10	LA_s+PTE_v	37.85
LA_s	39.80	LA_d+PTE_p	36.50
PTE_v	36.46	LA_w+PTE_p	36.10
PTE_p	36.57	LA_s+PTE_p	39.31

The OOV reduction rates (in %) and the total amount of OOV sentences/words of each method are given in Table 5. A large reduction can be seen when the methods LA_s and LA_w are applied to the original evaluation corpus, i.e., it is 6.2% (0.9%) for the lexical approximation with (without) skeleton matching. Both methods outperform the string similarity approach LA_d. The number of sentences with OOV words decreased by 26.3% for LA_s and the number of translated words increased, whereby the average sentence length of the obtained translations for sentences with recovered OOV words increased from 6.3 to 7.0 words per sentence. Concerning the automatic evaluation scores, only modest improvements could be achieved for LA_w and LA_d (BLEU: +0.5%), but large gains were obtained for LA_s (BLEU: +3.2%) when compared to the baseline system (cf. Table 7). Moreover, all improvements of the LA methods are significantly higher than the baseline results.

4.2 Effects of Phrase-Table Extension

The phrase-table generated from the Hindi-English training corpus contained 608,627 translation phrase-pairs, in which 11,457 source vocabulary words did not have a single-word entry. After filtering the phrase-table for the evaluation data set, the phrase-table contained 41,033 translation phrase-pairs. We applied two different phrase-table extension methods using (a) the intermediate Viterbi word alignments (PTE_v) as proposed in [7] and (b) the context of the original phrase-table (PTE_p) as proposed in this paper, adding a total of 1,133 new phrase-pairs to the original phrase-table.

The effects of the phrase-table extension are shown in Table 7. The only difference between the systems is the usage of the original phrase-table (*baseline*) versus the extended phrase-tables (PTE_v and PTE_p). All other decoder parameters were kept the same. The results show that the scores for PTE_v and PTE_p are slightly lower than the baseline results, although the differences are not statistically significant. Similar to dictionary-based phrase-table extension methods, the main reason for this effect is that no word information is available for the lexical phrase-pairs added by the PTE method. Therefore, ngram-based automatic evaluation metrics like BLEU that are highly sensitive to word order errors result in lower scores. However, compared to the baseline system, both added phrase-table entries increase the coverage of words in the evaluation data, thus reducing the number of OOV words in the baseline system output by 2.6% (cf. Table 6).

Comparing both phrase-table extension methods, slightly better results were achieved for the PTE_p compared to the PTE_v method. An examination revealed that 82.8% of the target phrases selected by both methods were identical. A paired comparison of the remaining translation phrase-pairs was carried out by a bilingual native speaker of Hindi and the results showed that 22.3% of the PTE_p target phrases were judged to have a superior translation.

4.3 Combination of LA and PTE

In order to combine both methods, we applied the lexical approximation methods to replace OOV words with appropriate variant word forms in the input and used the extended phrase-table (PTE_p) during SMT decoding as illustrated in Fig. 1.

4.3.1 Automatic Evaluation

The automatic evaluation scores are summarized in Table 7. The results show that the combined methods using skeleton matching largely outperform the baseline system (BLEU: up to +2.7%) whereby the improvements are statistically significant. On the other hand, the phrase-table extension still leads to lower BLEU scores when combined with the LA_w and LA_d methods whereby the combined methods using PTE_p outperforms the one using PTE_v. However, their differences towards the baseline system are not statistically significant. Concerning the coverage of translatable words, the combination of the *LA* and *PTE* methods further reduce the percentage of OOV words in the translation output. In total, only 24.9% of the evaluation data translation outputs still contain an untranslatable word after the LA_s+PTE_p method is applied, compared to 62.0% of the baseline system translations. The number of OOV words is reduced by 8.0% from 11.4% (baseline: 471 OOVs) to 3.4% (LA_s+PTE_p: 140 OOVs).

4.3.2 Subjective Evaluation

The automatic evaluation metrics are designed to judge the translation quality of the MT system outputs on a document level, i.e., scores are calculated on the sets of all evaluation data sentences, but not at the sentence level. In order

Table 8 Subjective evaluation. (Ranking)

baseline vs.	RANK	GAIN	Better	Same	Worse
baseline	2.68	—	—	—	—
PTE_p	2.69	+ 0.5%	0.8%	98.9%	0.3%
LA_w	2.77	+ 4.8%	6.4%	92.0%	1.6%
LA_s	3.39	+39.5%	50.2%	39.1%	10.7%
LA_s+PTE_p	3.47	+42.7%	57.9%	26.9%	15.2%

Table 9 Paired comparison evaluation ranks.

Better:	The rank of the proposed system output is better than the rank of the baseline system.
Same:	Both systems were assigned the same rank.
Worse:	The rank of the proposed system output is worse than the rank of the baseline system.

to get an idea of how much the translation quality of a single sentence is effected by the proposed methods, a subjective evaluation using the *ranking* metric is applied, whereby the *baseline* system is compared to the outputs of both lexical approximation methods (LA_w, LA_s), the phrase-table extraction method (PTE_p) and the combination of lexical approximation with skeleton matching and phrase-table extension (LA_s+PTE_p). For human assessment, all input sentences containing OOV words that could not be addressed by these methods were ignored. In total, 250 sentences were used for the evaluation of the above systems. Table 8 gives the normalized ranking results of three human evaluators for the Hindi-English translation task. Based on the ranking evaluation, each system output was compared against the baseline system using the pairwise-comparison criteria listed in Table 9. The *gain* of the proposed methods towards the baseline is calculated as the difference in the percentages of improved and degraded translations (*%better* - *%worse*).

The results show that all proposed methods are ranked better than the baseline system, although the differences between the phrase-table extension method (PTE_p) and the baseline system are not statistically significant. As indicated by the automatic evaluation scores, the lexical approximation methods achieve much better translations, especially when the skeleton match is applied. Interestingly, the combination of the *LA* and the *PTE* methods is ranked higher than the lexical approximation method alone, which indicates that, despite slightly lower automatic evaluation results, the phrase-table extension method actually helps to boost system performance.

4.3.3 OOV Translation Quality

In addition to *ranking*, a subjective evaluation of the appropriateness of the lexical approximation as well as of the quality of the OOV word translations was carried out by one native speaker of Hindi. For each OOV word in the MT output of the best performing system according to the ranking evaluation (LA_s+PTE_p), the respective sentences (original input and modified input after the lexical approximation), the OOV word and its lexical approximation, as well as the translation output, were given to the evaluator who had to judge whether (a) the lexical approximation and (b) the translation of the OOV word were *correct* or *incorrect*. The results summarized in Table 10 show that good lexical approximations were achieved for 75% of the OOV words, so that 65% of the OOV words were translated correctly by the LA_s+PTE_p method.

Table 10 OOV translation quality.

LA_s+PTE_p	Lexical Approximation	Translation Accuracy
correct	75.2%	65.4%
incorrect	24.8%	34.6%

Table 11 Translation examples.

input:	kyA mEM goluPa korsa kA ArakRaNa mila sakawA hE? (*Could I make a reservation for the golf course?*)
(OOV)	"goluPa"→ [LA] "golf"
baseline:	Can I have a reservation course?
LA_s+PTE_p:	Can I have a reservation for the golf course?
input:	kripyA muJe kuCa Ora xusare rango me xiKAo . (*Please show me some others in different colors.*)
(OOV)	"rango"→ [LA] "color" "xiKAo"→ [LA] "show"
baseline:	I 'd like something in second, please.
LA_s+PTE_p:	Will you show me some others in another color.
input:	ye kamarA mere lie bahuwa meMhagA hE . (*This room is too expensive for me.*)
(OOV)	"meMhagA"→ [LA] "expensive"
baseline:	This room is too long for me.
LA_s+PTE_p:	This room is too expensive for me.

4.3.4 Translation Examples

Table 11 provides some examples of the subjective evaluation results where the LA_s+PTE_p method was successfully applied, resulting in a better translation than the baseline system output. In most cases, the lexical approximation method modifies the unknown word to an expression that exists in the phrase-table, so that the context of the newly obtained source language phrase enables the SMT decoder to select appropriate translation equivalences. For example, the unknown word "*goluPa*" is approximated with "*golaPa*" that triggers the selection of the phrase-pair "*golaPa korsa ||| golf course*" during decoding.

4.3.5 Target Language Dependency

In addition to English, the methods proposed in this paper, were also evaluated for translation tasks using Chinese (C) and Japanese (J) as the target language. The automatic evaluation results are given in Table 12 and confirm the difficulty of the translation tasks indicated by the language perplexity figures given in Table 4. In general, the findings concerning the Hindi-English translation task given in the previous sections are confirmed for Chinese and Japanese as well, although the gains are much smaller.

Table 12 Target language dependency evaluation.

BLEU%	E	C	J
baseline	36.59	24.03	39.72
LA_d	37.05	24.04	40.14
LA_w	37.10	24.04	40.35
LA_s	39.80	24.52	40.93

BLEU%	E	C	J
PTE_v	36.46	23.23	39.31
PTE_p	36.57	23.97	39.77
LA_w+PTE_p	36.10	23.91	39.83
LA_s+PTE_p	39.31	23.53	40.14

5. 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-English, Hindi-to-Chinese, and Hindi-to-Japanese revealed that the combination of both methods significantly improved the translation quality for input sentences containing OOV words, although there are differences in gains according to the amount of training resources as well as the target language. 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, which should boost the overall performance of the proposed method further.

Acknowledgements

This work is partly supported by the Grant-in-Aid for Scientific Research (C) Number 19500137 and "Construction of speech translation foundation aiming to overcome the barrier between Asian languages", the Special Coordination Funds for Promoting Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology, Japan. The authors would also like to thank G. Varkey, V.N. Shukla, and S.S. Agrawal of CDAC Noida and S. Nakamura of NICT for constant support and conducive environment for this work.

References

[1] P. Koehn, F. Och, and D. Marcu, "Statistical phrase-based translation," Proc. HLT-NAACL, pp.127–133, Edmonton, Canada, 2003.
[2] D. Chiang, "A hierarchical phrase-based model for statistical machine translation," Proc. 43rd ACL, pp.263–270, Ann Arbor, Michigan, 2005.
[3] H. Okuma, H. Yamamoto, and E. Sumita, "Introducing Translation Dictionary into phrase-based SMT," Proc. MT Summit XI, pp.361–368, Copenhagen, Denmark, 2007.
[4] M. Popovic and H. Ney, "Exploiting phrasal lexica and additional morpho-synatctic language resources for SMT with scarce training data," Proc. EAMT, pp.212–218, Budapest, Hungary, 2005.
[5] P. Koehn and H. Hoang, "Factored translation models," Proc. EMNLP-CoNLL, pp.868–876, Prague, Czech Republic, 2007.
[6] K. Knight and J. Graehl, "Machine transliteration," Proc. 35th ACL, pp.128–135, Madrid, Spain, 1997.
[7] C. Mermer, H. Kaya, Ö. Güneş, and M. Doğan, "The TÜBİTAK-UEKAE SMT System for IWSLT 2008," Proc. IWSLT, pp.138–142, Hawaii, USA, 2008.
[8] WX, "Roman transliteration scheme for Devanagar." http://sanskrit.inria.fr/DATA/wx.html, 2007.
[9] P. Koehn, H. Hoang, A. Birch, and C. Callison-Burch, "Moses: Open source toolkit for SMT," Proc. 45th ACL, pp.177–180, Prague, Czech Republic, 2007.
[10] G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita, "Comparative study on corpora for speech translation," IEEE Trans. Audio Speech Language, vol.14, no.5, pp.1674–1682, 2006.
[11] F. Och and H. Ney, "A systematic comparison of various statistical alignment models," Computational Linguistics, vol.29, no.1, pp.19–51, 2003.
[12] A. Stolcke, "SRILM - An extensible language modeling toolkit," Proc. ICSLP, pp.901–904, Denver, USA, 2002.
[13] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of MT," Proc. 40th ACL, pp.311–318, Philadelphia, USA, 2002.
[14] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "Further meta-evaluation of machine translation," Proc. 3rd Workshop on SMT, ACL, pp.70–106, Columbus, Ohio, 2008.
[15] Y. Zhang, S. Vogel, and A. Waibel, "Interpreting Bleu/NIST scores: How much improvement do we need to have a better system?," Proc. LREC, pp.2051–2054, Lisbon, Portugal, 2004.

Michael Paul received his M.S. degree in computer science from the University of Saarland, Germany in 1994 and Ph.D. degree in engineering from Kobe University, Japan in 2006. He is currently a researcher in the MASTAR Project, NICT. His research interests include machine translation, evaluation of translation quality, and machine learning.

Karunesh Arora received his B.Tech. degree in computer engineering from NIT Surathkal, India in 1991. He is currently a research scientist with CDAC Noida, India. His research interests include machine translation, cross lingual information access and speech database development.

Eiichiro Sumita received his M.S. degree in computer science from the University of Electro-Communications in 1982 and Ph.D. degree in engineering from Kyoto University in 1999. He is the leader of NICT-KCCRC-LTG, a visiting professor of Kobe University. His research interests include machine translation and e-Learning.