Monolingual Distributional Profiles for Word Substitution in Machine Translation

Rashmi Gangadharaiah rgangadh@cs.cmu.edu
Ralf D. Brown ralf@cs.cmu.edu
Jaime Carbonell jgc@cs.cmu.edu

Language Technologies Institute,
Carnegie Mellon University

Abstract

Out-of-vocabulary (OOV) words present a significant challenge for Machine Translation. For low-resource languages, limited training data increases the frequency of OOV words and this degrades the quality of the translations. Past approaches have suggested using stems or synonyms for OOV words. Unlike the previous methods, we show how to handle not just the OOV words but *rare* words as well in an Example-based Machine Translation (EBMT) paradigm. Presence of OOV words and rare words in the input sentence prevents the system from finding longer phrasal matches and produces low quality translations due to less reliable language model estimates. The proposed method requires only a monolingual corpus of the source language to find candidate replacements. A new framework is introduced to score and rank the replacements by efficiently combining features extracted for the candidate replacements. A lattice representation scheme allows the decoder to select from a beam of possible replacement candidates. The new framework gives statistically significant improvements in English-Chinese and English-Haitian translation systems.

1 Introduction

An EBMT system makes use of a parallel corpus to translate new sentences. Each input sentence is matched against the source side of a training corpus. When matches are found, the corresponding translations in the target language are obtained through sub-sentential alignment. In our EBMT system, the final translation is obtained by combining the partial target translations using a statistical target Language Model. EBMT systems, like other data-driven approaches, require large amounts of data to function well (Brown, 2000).

Having more training data is beneficial resulting in log-linear improvement in translation quality for corpus-based methods (EBMT, SMT). Koehn (2002) shows translation scores for a number of language pairs with different training sizes translated using the Pharaoh SMT toolkit (Koehn et al., 2003). However, obtaining sizable parallel corpora for many languages is time-consuming and expensive. For rare languages, finding bilingual speakers becomes especially difficult.

One of the main reasons for low quality translations is the presence of large number of OOV and rare words (low frequency words in the training corpus). Variation in domain and errors in spelling increase the number of OOV words. Many of the present translation systems either ignore these unknown words or leave them untranslated in the final target translation. When data is limited, the number of OOV words increases, leading to the poor performance of the translation models and the language models due to the absence of longer sequences of source word matches and less reliable language model estimates.

Approaches in the past have suggested using stems or synonyms for OOV words as replacements (Yang and Kirchhoff, 2006). Similarity measures have been used to find words that are closely related (Marton et al., 2009). For morpho-

Coling 2008: Poster Volume, pages 320–328,
Beijing, August 2010

logically rich languages, the OOV word is morphologically analyzed and the stem is used as its replacement (Popović and Ney, 2004).

This paper presents a simpler method inspired by the Context-based MT approach (Carbonell et al., 2006) to improve translation quality. The method requires a large source language monolingual corpus and does not require any other language dependent resources to obtain replacements. Approaches suggested in the past only concentrated on finding replacements for the OOV words and not the rare words. This paper proposes a unified method to find possible replacements for OOV words as well as rare words based on the context in which these words appear. In the case of rare words, the translated sentence is traced back to find the origin of the translations and the target translations of the replacements are replaced with the translations of the rare words. In the case of OOV words, the target translations are replaced by the OOV word itself. The main idea for adopting this approach is the belief that the EBMT system will be able to find longer phrasal matches and that the language model will be able to give better probability estimates while decoding if it is not forced to fragment text at OOV and rare-word boundaries. This method is highly beneficial for low-resource languages that do not have morphological analysers or Part-of-Speech (POS) taggers and in cases where the similarity measures proposed in the past do not find closely related words for certain OOV words.

The rest of the paper is organized as follows. The next section (Section 2) discusses related work in handling OOV words. Section 3 describes the method adopted in this paper. Section 4 describes the experimental setup. Section 5 reports the results obtained with the new framework for English-Chinese and English-Haitian translation systems. Section 6 concludes and suggests possible future work.

2 Related Work

Orthographic and morpho-syntactic techniques for preprocessing training and test data have been shown to reduce OOV word rates. Popović and Ney (2004) demonstrated this on rich morphological languages in an SMT system. They introduced different types of transformations to the verbs to reduce the number of unseen word forms. Habash (2008) addresses spelling, name-transliteration OOVs and morphological OOVs in an Arabic-English Machine Translation system. Phrases with the OOV replacements in the phrase table of a phrase-based SMT system were "recycled" to create new phrases in which the replacements were replaced by the OOV words.

Yang and Kirchhoff (2006) proposed a back-off model for phrase-based SMT that translated word forms in the source language by hierarchical morphological phrase level abstractions. If an unknown word was found, the word was first stemmed and the phrase table entries for words sharing the same stem were modified by replacing the words with their stems. If a phrase entry or a single word phrase was found, the corresponding translation was used, otherwise the model backed off to the next level and applied compound splitting to the unknown word. The phrase table included phrasal entries based on full word forms as well as stemmed and split counterparts.

Vilar et al. (2007) performed the translation process treating both the source and target sentences as a string of letters. Hence, there are no unknown words when carrying out the actual translation of a test corpus. The word-based system did most of the translation work and the letter-based system translated the OOV words.

The method proposed in this work to handle OOV and rare words is very similar to the method adopted by Carbonell et al. (2006) to generate word and phrasal synonyms in their Context-based MT system. Context-based MT does not require parallel text but requires a large monolingual target language corpus and a fullform bilingual dictionary. The main principle is to find those n-gram candidate translations from a large target corpus that contain as many potential word and phrase translations of the source text from the dictionary and fewer spurious content words. The overlap decoder combines the target n-gram translation candidates by finding maximal left and right overlaps with the translation candidates of the previous and following n-grams. When the overlap decoder does not find coherent sequences of overlapping target n-grams, more candidate transla-

tions are obtained by substituting words or phrases in the target n-grams by their synonyms.

Barzilay and McKeown (2001) and Callison-Burch et al. (2006) extracted paraphrases from monolingual parallel corpus where multiple translations were present for the same source. The synonym generation in Carbonell et al. (2006) differs from the above in that it does not require parallel resources containing multiple translations for the same source language. In Carbonell et al. (2006), a list of paired left and right contexts that contain the desired word or phrase are extracted from the monolingual corpus. The same corpus is used to find other words and phrases that fit the paired contexts in the list. The idea is based on the distributional hypothesis which states that words with similar meanings tend to appear in similar contexts (Harris, 1954). Hence, their approach performed synonym generation on the target language to find translation candidates that would provide maximal overlap during decoding.

Marton et al. (2009) proposed an approach similar to Carbonell et al. (2006) to obtain replacements for OOV words, where monolingual distributional profiles for OOV words were constructed. Hence, the approach was applied on the source language side as opposed to Carbonell et al. (2006) which worked on the target language. Only similarity scores and no other features were used to rank the paraphrases (or replacements) that occured in similar contexts. The high ranking paraphrases were used to augment the phrase table of phrase-based SMT.

All of the previously suggested methods only handle OOV words (except Carbonell et al. (2006) which handles low frequency target phrases) and no attempt is made to handle rare words. Many of the methods explained above directly modify the training corpus (or phrase table in phrase-based SMT) increasing the size of the corpus. Our method clusters words and phrases based on their context as described by Carbonell et al. (2006) but uses the clustered words as replacements for not just the OOV words but also for the rare words on the source language side. Our method does not make use of any morphological analysers, POS taggers or manually created dictionaries as they may not be available for many rare or low-resource languages. The translation of the replacements in the final decoded target sentence is replaced by the translation of the original word (or the source word itself in the OOV case), hence, we do not specifically look for synonyms. The only condition for a word to be a candidate replacement is that its left and right context need to match with that of the OOV/rare-word. Hence, the clustered words could have different semantic relations. For example,

$(cluster1)$:"laugh, giggle, chuckle, cry, weep"
where "laugh, giggle, chuckle" are synonyms and "cry, weep" are antonyms of "laugh".

Clusters can also contain hypernyms (or hyponyms), meronyms (or holonyms), troponyms and coordinate terms along with synonyms and antonyms. For example,

$(cluster2)$:"country, region, place, area, district, state, zone, United States, Canada, Korea, Malaysia".
where "country" is a hypernym of "United States/Canada/Korea/Malaysia". "district" is a meronym of "state". "United States, Canada, Korea, Malaysia" are coordinate terms sharing "country" as their hypernym.

The contributions made by the paper are threefold: first, replacements are found for not just the OOV words but for the *rare* words as well. Second, the framework used allows scoring replacements based on multiple features to permit optimization. Third, instead of directly modifying the training corpus by replacing the candidate replacements by the OOV words, a new representation scheme is used for the test sentences to efficiently handle a beam of possible replacements.

3 Proposed Method

Like Marton et al. (2009), only a large monolingual corpus is required to extract candidate replacements. To retrieve more replacements, the monolingual corpus is pre-processed by first generalizing numbers, months and years by NUMBER, MONTH and YEAR tags, respectively.

3.1 OOV and Rare words

Words in the test sentence (new source sentence to be translated) that do not appear in the training corpus are called OOV words. Words in the test sentence that appear less than K times in the training corpus are considered as rare words (in this paper $K = 3$). The method presented in the following sections holds for both OOV as well as rare words. In the case of rare words, the final translation is postprocessed (Section 3.7) to include the translation of the rare word.

The procedure adopted will be explained with a real example T (the rest of the sentence is removed for the sake of clarity) encountered in the test data with "hawks" as the OOV word,

T:a mobile base , *hitting three* **hawks** *with one arrow* over the past few years ...

3.2 Context

As the goal is to obtain longer target phrasal translations for the $test\ sentence$ before decoding, only words that fit the left and right context of the OOV/rare-word in the test sentence are extracted. Unlike Marton et al. (2009) where a context list for each OOV is generated from the contexts of their replacements, this paper uses only the left and right context of the OOV/rare-word. The default window size for the context is five words (two words to the left and two words to the right of the OOV/rare-word). If the windowed words contain only function words, the window is incremented until at least one content word is present in the resulting context. This enables one to find sensible replacements that fit the context well. The contexts for T are:

Left-context (L): hitting three
Right-context (R): with one arrow

The above contexts are further processed to generalize the numbers by a $NUMBER$ tag to produce more candidate replacements. The resulting contexts are now:

Left-context (L): hitting $NUMBER$
Right-context (R): with $NUMBER$ arrow

As a single $L - R$ context is used, a far smaller number of replacements are extracted.

3.3 Finding Candidate replacements

The monolingual corpus (ML) of the source language is used to find words and phrases (X_k) that fit LX_kR i.e., with L as its left context and/or R as its right context. The maximum length for X_k is set to 3 currently. The replacements are further filtered to obtain only those replacements that contain at least one content word. As illustrated earlier, the resulting replacement candidates are not necessarily synonyms.

3.4 Features

A local context of two to three words to the left of an OOV/rare-word ($word_i$) and two to three words to the right of $word_i$ contain sufficient clues for the word, $word_i$. Hence, local contextual features are used to score each of the replacement candidates ($X_{i,k}$) of $word_i$. Each $X_{i,k}$ extracted in the previous step is converted to a feature vector containing 11 contextual features. Certainly more features can be extracted with additional knowledge sources. The framework allows adding more features, but for the present results, only these 11 features were used.

As our aim is to assist the translation system in finding longer target phrasal matches, the features are constructed from the occurrence statistics of $X_{i,k}$ from the bilingual training corpus (BL). If a candidate replacement does not occur in the BL, then it is removed from the list of possible replacement candidates.

Frequency counts for the features of a particular replacement, $X_{i,k}$, extracted in the context of $L_{i,-2}L_{i,-1}$ (two preceding words of $word_i$) and $R_{i,+1}R_{i,+2}$ (two following words of $word_i$) (the remaining words in the left and right context of $word_i$ are not used for feature extraction) are obtained as follows:

f_1: frequency of $X_{i,k}R_{i,+1}$
f_2: frequency of $L_{i,-1}X_{i,k}$
f_3: frequency of $L_{i,-1}X_{i,k}R_{i,+1}$
f_4: frequency of $L_{i,-2}L_{i,-1}X_{i,k}$
f_5: frequency of $X_{i,k}R_{i,+1}R_{i,+2}$
f_6: frequency of $L_{i,-2}L_{i,-1}X_{i,k}R_{i,+1}$

f_7: frequency of $L_{i,-1}X_{i,k}R_{i,+1}R_{i,+2}$
f_8: frequency of $L_{i,-2}L_{i,-1}X_{i,k}R_{i,+1}R_{i,+2}$
f_9: frequency of $X_{i,k}$ in ML
f_{10}: frequency of $X_{i,k}$ in BL
f_{11}: number of feature values $(f_1, ..f_{10}) > 0$

f_{11} is a vote feature which counts the number of features (f_1 ... f_{10}) that have a value greater than zero. The features are normalized to fall within $[0, 1]$. The sentences in ML, BL and test data are padded with two begin markers and two end markers for obtaining counts for OOV/rare-words that appear at the beginning or end of a test sentence.

3.5 Representation

Before we go on to explaining the lattice representation, we would like to make a small clarification in the terminalogy used. In the MT community, a lattice usually refers to the list of possible partially-overlapping target translations for each possible source $n-$gram phrase in the input sentence. Since we are using the term lattice to also refer to the possible paths through the input sentence, we will call the lattice used by the decoder, the "*decoding lattice*". The lattice obtained from the input sentence representing possible replacement candidates will be called the "*input lattice*".

An input lattice (Figure 1) is constructed with a beam of replacements for the OOV and rare words. Each replacement candidate is given a score (Eqn 1) indicating the confidence that a suitable replacement is found. The numbers in Figure 1 indicate the start and end indices (based on character counts) of the words in the test sentence. In T, two replacements were found for the word *"hawks"*: *"homers"* and *"birds"*. However, "homers" was not found in the BL and hence, it was removed from the replacement list.

The input lattice also includes the OOV word with a low score (Eqn 2). This allows the EBMT system to also include the OOV/rare-word during decoding. In the Translation Model of the EBMT system, this test lattice is matched against the source sentences in the bilingual training corpus. The matching process would now also look for phrases with "birds" and not just "hawks". When a match is found, the corresponding translation in the target language is obtained through sub-sentential alignment (Section 3.7). The scores on the input lattice are later used by the decoder (Section 3.7). Each replacement $X_{i,k}$ for the OOV/rare-word ($word_i$) is scored with a logistic function (Bishop, 2006) to convert the dot product of the features and weights ($\vec{\lambda} \cdot \vec{f_{i,k}}$) to a score between 0 and 1 (Eqn 1 and Eqn 2).

T : a mobile base , hitting three hawks with one arrow .....

input lattice:

0	0	(" a ")
1	6	(" mobile ")
7	10	(" base ")
11	11	(",")
12	18	(" hitting ")
13	17	(" three ")
18	22	(" *hawks* " 0.0026)
18	22	(" **birds** " 0.9974)
23	26	(" with ")
27	29	(" one ")
30	34	(" arrow ")

Figure 1: Lattice of the input sentence T containing replacements for OOV words.

OOV/Rare word	Candidate Replacements
Spelling errors	
krygyzstan	kyrgyzstan,...
yusukuni	yasukuni,..
kilomaters	kilometers, miles, km, ...
	Coordinate terms
somoa	india, turkey, germany, russia, japan,...
ear	body, arms, hands, feet, mind, car, ...
buyers	dealer, inspector, the experts, smuggler,.
	Synonyms
plummet	drop, dropped, fell, ...
	Synonyms and Antonyms
optimal	worse, better, minimal,....

Figure 2: Sample English candidate replacements obtained.

$$p_\lambda(X_{i,k}|word_i) = \frac{exp(\vec{\lambda}\vec{f_{i,k}})}{1+\sum_{j=1...S} exp(\vec{\lambda}\vec{f_{i,j}})} \quad (1)$$

$$p_\lambda(word_i) = \frac{1}{1 + \sum_{j=1...S} exp(\vec{\lambda} \cdot \vec{f_{i,j}})} \quad (2)$$

where, $\vec{f_{i,j}}$ is the feature vector for the j^{th} replacement candidate of $word_i$, S is the number of replacements, $\vec{\lambda}$ is the weight vector indicating the importance of the corresponding features.

3.6 Tuning feature weights

We would like to select those feature weights ($\vec{\lambda}$) which would lead to the least expected loss in translation quality (Eqn 3). $-log(BLEU)$ (Papineni et al., 2002) is used to calculate the expected loss over a development set. As this objective function has many local minima and is piecewise constant, the surface is smoothed using the L2-norm regularization. Powell's algorithm (Powell, 1964) with grid-based line optimization is used to find the best weights. 7 different random guesses are used to initialize the algorithm.

$$\min_{\lambda} E_\lambda[L(t_{tune})] + \tau * ||\lambda||^2 \quad (3)$$

The algorithm assumes that partial derivates of the function are not available. Approximations of the weights ($\lambda_1, ..\lambda_N$) are generated successively along each of the N standard base vectors. The procedure is iterated with a stopping criteria based on the amount of change in the weights and the change in the loss. A cross-validation set (in addition to the regularization term) is used to prevent overfitting at the end of each iteration of the Powell's algorithm. This process is repeated with different values of τ, as in Deterministic Annealing (Rose, 1998). τ is initialized with a high value and is halved after each process.

3.7 System Description

The EBMT system finds phrasal matches for the test (or input) sentence from the source side of the bilingual corpus. The corresponding target phrasal translations are obtained through subsentential alignment. When an *input lattice* is given instead of an input sentence, the system performs the same matching process for all possible phrases obtained from the input lattice. Hence, the system also finds matches for source phrases that contain the replacements for the OOV/rare-word. Only the top C ranking replacement candidates for every OOV/rare word are used in building the input lattice. The optimal value of C was empirically found to be 2. On examining the obtained input lattices, the proposed method found replacements for at the most 3 OOV/rare words in each test sentence (Section 4). Hence, the number of possible paths through the input lattice is not substantially large.

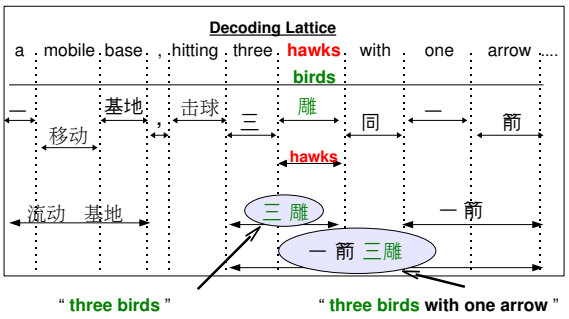

Figure 3: Lattice containing possible phrasal target translations for the test sentence T.

The target translations of all the source phrases are placed on a common decoding lattice. An example of a decoding lattice for example T is given in Figure 3. The system is now able to find longer matches (*" three birds with one arrow "* and *" three birds "*) which was not possible earlier with the OOV word, *"hawks"*. The local ordering information between the translations of "three birds" and "with one arrow" is well captured due to the retrieval of the longer source phrasal match, "three birds with one arrow". Our ultimate goal is to obtain translations for such longer $n-$gram source phrases boosting the confidence of both the translation model and the language model.

The decoder used in this paper (Brown, 2003) works on this *decoding lattice* of possible phrasal target translations (or fragments) for source phrases present in the *input lattice* to generate the target translation. Similar to Pharaoh (Koehn et al., 2003), the decoder uses multi-level beam search with a priority queue formed based on the number of source words translated. Bonuses are given for paths that have overlapping fragments. The total score (TS) for a path (Eqn 4) through the translation lattice is the arithmetic average of the scores for each target word in the

path. The EBMT engine assigns each candidate phrasal translation a quality score computed as a log-linear combination of alignment score and translation probability. The alignment score indicates the engine's confidence that the right target translation has been chosen for a source phrase. The translation probability is the proportion of times each distinct alternative translation was encountered out of all the translations. If the path includes a candidate replacement, the log of the score, $p_\lambda(w_i)$, given for a candidate replacement is incorporated into TS as an additional term with a weight wt_5.

$$TS = \frac{1}{t}\sum_{i=1}^{t}[wt_1 \log(b_i) + wt_2 \log(pen_i) + wt_3 \log(q_i) + wt_4 \log(P(w_i|w_{i-2}, w_{i-1})) + \mathbb{1}_{(w_i=replacement)} wt_5 \log(p_\lambda(w_i)) \ ] \quad (4)$$

where, t is the number of target words in the path, wt_j indicates the importance of each score, b_i is the bonus factor given for long phrasal matches, pen_i is the penalty factor for source and target phrasal-length mismatches, q_i is the quality score and $P(w_i|w_{i-2}, w_{i-1})$ is the LM score. The parameters of the EBMT system (wt_j) are tuned on a development set.

The target translation is postprocessed to include the translation of the OOV/rare-word with the help of the best path information from the decoder. In the case of OOV words, since the translation is not available, the OOV word is put back into the final output translation in place of the translation of its replacement. In the output translation of the test example T, the translation of *"birds"* is replaced by the word, *"hawks"*. For rare words, knowing that the translation of the rare word may not be correct (due to poor alignment statistics), the target translation of the replacement is replaced by the translation of the rare word obtained from the dictionary. If the rare word has multiple translations, the translation with the highest score is chosen.

4 Experimental Setup

As we are interested in improving the performance of low-resource EBMT, the English-Haitian (Eng-Hai) newswire data (Haitian Creole, CMU, 2010) containing 15,136 sentence-pairs was used. To test the performance in other languages, we simulated sparsity by choosing less training data for English-Chinese (Eng-Chi). For the Eng-Chi experiments, we extracted 30k training sentence pairs from the FBIS (NIST, 2003) corpus. The data was segmented using the Stanford segmenter (Tseng et al., 2005). Although we are only interested in small data sets, we also performed experiments with a larger data set of 200k. 5-gram Language Models were built from the target half of the training data with Kneser-Ney smoothing. For the monolingual English corpus, 9 million sentences were collected from the Hansard Corpus (LDC, 1997) and FBIS data.

EBMT system without OOV/rare-word handling is chosen as the Baseline system. The parameters of the EBMT system are tuned with 200 sentence pairs for both Eng-Chi and Eng-Hai. The tuned EBMT parameters are used for the Baseline system and the system with OOV/rare-word handling. The feature weights for the proposed method are then tuned on a seperate development set of 200 sentence-pairs with source sentences containing at least 1 OOV/rare-word. The cross-validation set for this purpose is made up of 100 sentence-pairs. In the OOV case, 500 sentence pairs containing at least 1 OOV word are used for testing. For the rare word handling experiments, 500 sentence pairs containing at least 1 rare word are used for testing.

To assess the translation quality, 4-gram word-based BLEU is used for Eng-Hai and 3-gram word-based BLEU is used for Eng-Chi. Since BLEU scores have a few limitations, the NIST and TER metrics are also used. The test data used for comparing the system handling OOV words and the Baseline (without OOV word handling) is different from the test data used for comparing the system handling rare words and the Baseline system (without rare word handling). In the former case, the test data handles only OOV words and in the latter, the test data only handles rare words. Hence, the test data for both the cases do not completely overlap. As we are interested in determining whether handling rare words in test sentences is useful, we keep both the test data sets seperate and assess the improvements obtained by only

OOV/Rare	system	TER	BLEU	NIST
OOV	Baseline	77.89	18.61	4.8525
	Handling OOV	76.95	19.32	4.9664
Rare	Baseline	74.23	22.84	5.3803
	Handling Rare	74.02	23.12	5.4406

Table 1: Comparison of translation scores of the Baseline system and system handling OOV and Rare words for Eng-Hai.

handling OOV words and by only handling rare words over their corresponding Baselines. As future work, it would be interesting to create one test data set to handle both OOV and rare words to see the overall gain.

The test set is further split into 5 files and the Wilcoxon (Wilcoxon, 1945) Signed-Rank test is used to find the statistical significance.

5 Results

Sample replacements found are given in Figure 2. For both Eng-Chi and Eng-Hai experiments, only the top C ranking replacement candidates were used. The value of C was tuned on the development set and the optimal value was found to be 2. Translation quality scores obtained on the test data with 30k and 200k Eng-Chi training data sets are given in Table 2. Table 1 shows the results obtained on Eng-Hai. Statistically significant improvements ($p < 0.0001$) were seen by handling OOV words as well as rare words over their corresponding baselines.

As the goal of the approach was to obtain longer target phrasal matches, we counted the number of n-grams for each value of n present on the decoding lattice in the 30k Eng-Chi case. The subplots: A and B in Figure 4, shows the frequency of n-grams for higher values of n (for $n > 5$) when handling OOV and rare words. The plots clearly show the increase in number of longer target phrases when compared to the phrases obtained by the baseline systems.

Since the BLEU and NIST scores were computed only up to 3-grams, we further found the number of n-gram matches (for $n > 3$) in the final translation of the test data with respect to the reference translations (subplots: C and D). As expected, a larger number of longer $n-$gram matches were found. For the OOV case, matches

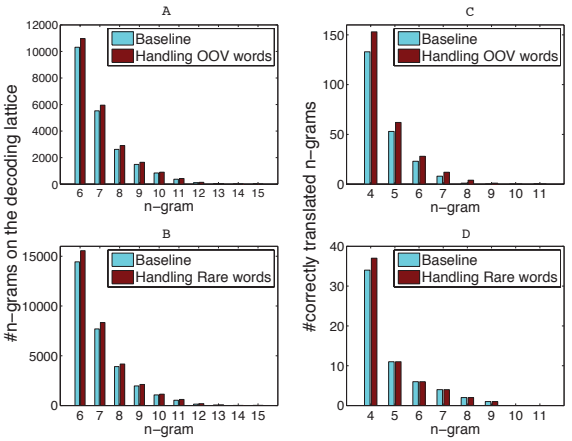

Figure 4: A, B: number of n-grams found for increasing values of n on the decoding lattice. C, D: number of target n-gram matches for increasing values of n with respect to the reference translations.

OOV/Rare	Training data size	system	TER	BLEU	NIST
OOV	30k	Baseline	82.03	14.12	4.1186
	30k	Handling OOV	80.97	14.78	4.1798
	200k	Baseline	79.41	19.90	4.6822
	200k	Handling OOV	77.66	20.50	4.7654
Rare	30k	Baseline	82.09	15.36	4.3626
	30k	Handling Rare	80.02	16.03	4.4314
	200k	Baseline	78.04	20.96	4.9647
	200k	Handling Rare	77.35	21.17	5.0122

Table 2: Comparison of translation scores of the Baseline system and system handling OOV and Rare words for Eng-Chi.

up to 9-grams were found where the baseline only found matches up to 8-grams.

6 Conclusion and Future Work

A simple approach to improve translation quality by handling both OOV and rare words was proposed. The framework allowed scoring and ranking each replacement candidate efficiently.

The method was tested on two language pairs and statistically significant improvements were seen in both cases. The results showed that rare words also need to be handled to see improvements in translation quality.

In this paper, the proposed method was only applied on words, as future work we would like to extend it to OOV and rare-phrases as well.

References

R. Barzilay and K. McKeown 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computaional Linguistics,* pp. 50-57.

C. M. Bishop 2006. Pattern Recognition and Machine Learning, Springer.

R. D. Brown, R. Hutchinson, P. N. Bennett, J. G. Carbonell, P. Jansen. 2003. Reducing Boundary Friction Using Translation-Fragment Overlap. In *Proceedings of The Ninth Machine Translation Summit,* pp. 24-31.

R. D. Brown. 2000. Automated Generalization of Translation Examples. In *Proceedings of The International Conference on Computational Linguistics,* pp. 125-131.

C. Callison-Burch, P. Koehn and M. Osborne. 2006. Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of The North American Chapter of the Association for Computational Linguistics,* pp. 17-24.

J. Carbonell, S. Klien, D. Miller, M. Steinbaum, T. Grassiany and J. Frey. 2006. Context-Based Machine Translation Using Paraphrases. In *Proceedings of The Association for Machine Translation in the Americas,* pp. 8-12.

N. Habash. 2008. Four Techniques for Online Handling of Out-of-Vocabulary Words in Arabic-English Statistical Machine Translation. In *Proceedings of Association for Computational Linguistics-08: HLT,* pp. 57-60.

Public release of Haitian Creole language data by Carnegie Mellon, 2010. *http://www.speech.cs.cmu.edu/haitian/*

Z. Harris. 1954. Distributional structure. *Word,* 10(23): 146-162.

P. Koehn. 2004. Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. *The Association for Machine Translation*.

P. Koehn, F. J. Och and D. Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of HLT:The North American Chapter of the Association for Computational Linguistics*.

P. Koehn 2002 Europarl: A multilingual corpus for evaluation of machine translation. Unpublished, *http://www.isi.edu/koehn/publications/europarl/*

Linguistic Data Consortium. 1997 Hansard Corpus of Parallel English and French. Linguistic Data Consortium, December. *http://www.ldc.upenn.edu/*

Y. Marton, C. Callison-Burch and P. Resnik. 2009. Improved Statistical Machine Translation Using Monolingually-derived Paraphrases. In *Proceeding of The Empirical Methods in Natural Language Processing,* pp. 381-390.

NIST. 2003. Machine translation evaluation. *http://nist.gov/speech/tests/mt/*

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of The Association for Computational Linguistics*. pp. 311-318.

M. Popović and H. Ney. 2004. Towards the use of Word Stems and Suffixes for Statistical Machine Translation. In *Proceedings of The International Conference on Language Resources and Evaluation*.

M. J. D. Powell. 1964. An efficient method for finding the minimum of a function of several variables without calculating derivatives *Computer Journal*. Volume 7, pp. 152-162.

K. Rose. 1998. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. In *Proceedings of The Institute of Electrical and Electronics Engineers,* pp. 2210-2239.

H. Tseng, P. Chang, G. Andrew, D. Jurafsky and C. Manning. 2005. A Conditional Random Field Word Segmenter. *Fourth SIGHAN Workshop on Chinese Language Processing*.

D. Vilar, J. Peter, and H. Ney. 2007. Can we translate letters? In *Proceedings of Association Computational Linguistics Workshop on SMT,* pp. 33-39.

M. Yang and K. Kirchhoff. 2006. Phrase-based back-off models for machine translation of highly inflected languages. In *Proceedings of European Chapter of the ACL,* 41-48.

F. Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics, 1,* 80-83. tool: *http://faculty.vassar.edu/lowry/wilcoxon.html*