Towards Zero Unknown Word in Neural Machine Translation

Xiaoqing Li†, Jiajun Zhang‡, Chengqing Zong†‡
† National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences
‡ CAS Center for Excellence in Brain Science and Intelligence Technology
{xqli,jjzhang,cqzong}@nlpr.ia.ac.cn

Abstract
Neural Machine translation has shown promising results in recent years. In order to control the
computational complexity, NMT has to employ a small vocabulary, and massive rare words outside the
vocabulary are all replaced with a single unk symbol. Besides the inability to translate rare words, this
kind of simple approach leads to much increased ambiguity of the sentences since meaningless unks
break the structure of sentences, and thus hurts the translation and reordering of the in-vocabulary
words. To tackle this problem, we propose a novel substitution-translation-restoration method. In sub-
stitution step, the rare words in a testing sentence are replaced with similar in-vocabulary words
based on a similarity model learnt from monolingual data. In translation and restoration steps, the
sentence will be translated with a model trained on new bilingual data with rare words replaced,
and finally the translations of the replaced words will be substituted by that of original ones. Experiments
on Chinese-to-English translation demonstrate that our proposed method can achieve more
than 4 BLEU points over the attention-based NMT. When compared to the recently proposed method
handling rare words in NMT, our method can also obtain an improvement by nearly 3 BLEU points.

1 Introduction
Neural machine translation is a recently proposed approach to
MT and has shown competing results to conventional translation
methods [Kalchbrenner and Blunsom, 2013; Cho et al.,
2014; Sutskever et al., 2014]. In neural machine translation,
the source sentence is converted into vector representation by
a neural network called encoder, then another neural network
called decoder generate target sentence word by word based
on source representation and target history. This framework
has several advantages over conventional translation meth-
ods. First, it does not need any domain knowledge as required
by conventional methods to design features. Second, the dis-
tributed representation allows NMT model to generalize well
and produce novel translations for source words and phrases,
while the symbolic representation in conventional MT makes
it impossible to generate translations beyond the rule table
extracted from the bilingual corpus. Third, the memory con-
sumption of NMT model is also much smaller.

Despite these advantages, NMT models have a major draw-
back in handling rare words. In order to control the computa-
tional complexity, which grows proportional to target vocab-
ulary size, most NMT systems limit the vocabulary to con-
tain only 30k to 80k most frequent words in both the source
and target side and convert rare words into a single unk sym-
bol. An obvious problem of this approach is that NMT model
cannot learn the translation of rare words. In particular, if a
source word is outside the source vocabulary or its translation
is outside the target vocabulary, the model will not be able to
generate proper translation for this word during testing. An-
other problem is that masking rare words with meaningless unk
will increase the ambiguity of the sentence. This can be
illustrated by the following three sentences,

   a) Mike chases the pet with mottle
   b) Mike chases the pet with scooter
   c) Mike chases the pet with Sullivan

Assume all the last words in the three sentences are rare
words. The word 'mottle' in sentence 1 modifies the object
'pet', and both the word 'scooter' and 'Sullivan' in sentence
b) and c) modifies the predicate, but one describes the tool
and the other describes the companion. The translation of the
preposition 'with' and the word order will be quite different
when translating the three sentences into Chinese. If the last
words are replaced by the unk symbol, the three sentences
will be the same. As a result, the model can only generate the
translation by chance.

To solve the above problems, we propose a novel rare word
replacement method based on similarity. During training,
word alignment will first be induced from bilingual corpus.
And each aligned word pair which contains rare word either
on the source side or the target side will be replaced with
similar in-vocabulary words, where the similarity model is
learned from a large mono-lingual corpus. Then this new
bilingual corpus with rare words replaced will be used to train
a NMT model. During testing, the rare words in input sen-
tence will also be replaced with similar in-vocabulary words.

†source vocabulary size contributes less to computational com-
plexity, but knowing how to translate source word to target unk is
not helpful, so the source vocabulary size is also limited.
After translation, a post-processing step is adopted to recover the translation of rare words. Experiments on Chinese to English translation task show that more than 4 points in BLEU score can be gained with our approach over the baseline. And the gain is also much larger than a previously proposed replacement method [Luong et al., 2015b].

2 Neural Machine Translation and Impact of Rare Words

In this section, we first give a brief introduction to neural machine translation and explain why NMT model could not employ large vocabulary. Then we quantitatively analyze how rare words impact the performance of NMT.

2.1 Neural Machine Translation

Neural machine translation is conceptually simple: it models the translation probability of a source sentence \( s = (s_1, s_2, ..., s_m) \) into target sentence \( t = (t_1, t_2, ..., t_n) \) with a single neural network as follows,

\[
p(t|s) = \prod_{i=1}^{n} p(t_i|t_{<i}, s)
\]

where the conditional probability is often parameterized with the encoder-decoder framework. The encoder reads the source sentence and encodes it into a sequence of hidden states \( h = (h_1, h_2, ..., h_m) \):

\[
h_i = f(s_i, h_{i-1})
\]

Then the decoder generates the translation word by word based on the target hidden states \( z = (z_1, z_2, ..., z_n) \):

\[
p(t_i|t_{<i}, s) = \frac{1}{Z} \exp \{ q(t_i, t_{i-1}, z_i, c_i) \}
\]

where

\[
z_i = g(t_{i-1}, z_{i-1}, c_i)
\]

\[
c_i = r(z_{i-1}, h)
\]

In above formulations, \( f, q, g \) and \( r \) are non-linear transformations and varies in different systems.

The most time consuming step in the network is the calculation of the normalization constant \( Z \), which is computed as follows,

\[
Z = \sum_{t' \in IV} \exp \{ q(t', t_{i-1}, z_i, c_i) \}
\]

According to this equation, we need to iterate over all target in-vocabulary words to calculate a non-linear transformation for each, and then sum them up\(^2\). So the total computational complexity will grow almost proportional to the target vocabulary size. Considering that it usually takes days to weeks to train a NMT model on a large corpus with a vocabulary size of 30k to 80k, training with the whole target vocabulary is obviously infeasible. So addressing the problem for rare words is quite necessary for neural machine translation.

2.2 Impact of Rare Words

As discussed in the introduction part, rare words cause two problems for neural machine translation. First, NTM model cannot learn translations for rare words because they are all converted to \( unk \) in the training data. Second, rare words increase the ambiguity of the sentence, which increases the difficulty to translate and reorder the rest in-vocabulary words in the sentence.

To quantitatively check the impact of the two factors, we design the following experiment. We extract 5 groups of Chinese sentences with different number of rare words (0-4) from the NIST Chinese to English translation data set. Each group contains 50 sentences together with their reference translations. In order to rule out the influence of sentence length, all the sentences in the 5 groups are between 20 to 30 words. We use the same system to translate these sentences and the corresponding performances are shown in Figure 1 (red line). It is obvious more rare words lead to worse performance. To simulate the impact of missing translation for rare words, we randomly set 1-4 words to \( unk \) in the translation of sentences in group 0. The result is shown as blue line. It could be inferred that the remaining gap between the red line and the horizontal line (denoting the performance of group 0) is caused by the increased ambiguity. According to the figure, when there are only one rare word in the sentence, the performance drop is mainly caused by missing translations, but when there are more rare words, increased ambiguity also contributes a lot to the performance drop.

3 Replace Rare Words with Similar words

The analysis in the above section shows the importance of keeping the sentence structure complete. So we propose...
to replace rare words in training and testing data with in-vocabulary words similar to them. The data processing diagram is shown in Figure 2.

In the training phase, we first learn a similarity model from a monolingual corpus, which is used to evaluate the similarity between words. We also need to learn word level alignment for sentence pairs in the bilingual corpus. As a byproduct, a lexical translation table can be derived from the aligned bilingual corpus. In our experiments, we only reserve the translation with the highest probability for each word in the table. Then the aligned word pairs which contain rare words will be replaced with in-vocabulary words similar to them. Finally, a NMT model will be learned from the new bilingual corpus.

In the testing phase, the rare words in testing sentence will be first replaced with similar in-vocabulary words. Then the sentence after replacement will be translated by the NMT model obtained in the training phase. With the help of the lexical translation model, the translation of those rare words will be substituted back into the generated target sentence to obtain the final result.

There are three issues not explained in detail in the diagram, including i) which words in the bilingual corpus will be replaced? ii) how to evaluate similarity between words? iii) How to recover the translation for rare words during testing? The following parts in this section will answer these questions.

### 3.1 Words to Be Replaced

Different languages are not perfectly corresponded in word level. For example, English articles are usually omitted when translated into Chinese. And the city name New York is just one word in Chinese. Sometimes the correspondence is even at phrase or sentence level, such as the translation of idioms. In this paper, we only handle word pairs with one-to-one mapping and rare words aligned to null. According to whether the source and the target word is rare, there are five cases.

- **unk to unk**, both the source and target word in the aligned pair are rare words. In this case we will replace the source word with a similar in-vocabulary word and the target word with the translation of the similar word.
- **unk to common**, only the source word is rare. In this case we will keep the target word and replace the source word with the translation of the target word.
- common to **unk**, only the target word is rare. In this case we will keep the source word and replace the target word with the translation of the source word.
- common to common, no replacement in this case.
- **unk to null or null to unk**, source or target rare word is not aligned to any word. In this case we simply remove the rare word from the sentence.

### 3.2 Similarity Model

Distributed word representation has been shown powerful to capture syntactic and semantic information about words, and it is widely applied in various tasks [Turian et al., 2010]. We adopt it here to find the most similar word for a given word \(w\) as follows,

\[
w^* = \arg \max_{w' \in IV} \text{sim}(w, w')
\]

in which \(IV\) denotes the set of in-vocabulary words, and the function \(\text{sim}\) is the cosine similarity between two word vectors.

\[
\text{sim}(w, w') = \cos(\text{vec}(w), \text{vec}(w'))
\]

However, since the word vectors and the lexical translation table are learned automatically from data, they may lead to inappropriate alternative for original translation pairs. For example, the most similar word to the rare word ‘善款’ (donation) at the end of the following sentence is ‘筹募’ (raise), which is in fact a synonym to the second to last word ‘募捐’. As a result, this sentence will be ungrammatical after replacement because it will has two neighbouring predicates with the same meaning.

中国 红十字会 为 新疆 灾区 募集 善款

China Red-Cross for Xinjiang disaster-area raise donation

As another example, the similarity model find a synonym word ‘不和’ to the rare word ‘失和’ (discord) in the following sentence, but the lexical translation table gives it a wrong translation ‘divorce’.
The bilingual data to train the NMT model is selected from LDC, which contains about 0.6M sentence pairs. To avoid spending too much training time on long sentences, all sentences pairs longer than 50 words either on the source side or on the target side are discarded. The alignment information needed for replacement are obtained by the Berkeley Aligner [Liang et al., 2006] on the same bilingual data. We use the word2vec toolkit [Mikolov et al., 2013] to train word vectors on the monolingual data, which is the combination of the source side of the bilingual data and Chinese Gigaword Xiaohua portion. The Chinese bi-directional language model is trained with kenlm [Heafield et al., 2013] on the same monolingual data, while the English language model is trained on the combination of the target side of the bilingual data and the English Gigaword.

The NIST 03 dataset is chosen as the development set, which is used to monitoring the training process and decide the early stop condition. And the NIST 04 to 06 are used as our testing set.

4.2 Training Details
The hyperparameters used in our network are described as follows. We limit both the source and target vocabulary to 30k in our experiments. This number of hidden units is 1,000 for both the encoder and decoder. And the word embedding dimension is 500 for all source and target words. The parameters in the network are updated with the adadelta algorithm.

To train the word vectors on monolingual data, we set the embedding dimension to 100 and the window size to 5. And we use top 10 most similar words in the similarity model considering bilingual context in section 3.2.

4.3 Main Results
We compare our best system (the one with bilingual similarity) to the baseline without any replacement, and the system proposed in [Luong et al., 2015b], which only annotate target unk as unk-\(k\), in which \(k\) indicates which source word translates into current unknown word. In particular, if \(e_j\) in the target sentence is a rare word and it’s aligned to source word \(c_i\), then \(k\) will be \(i-j\). The performance of Moses [Koehn et al., 2007] with 4-gram language model trained on the target side of the bilingual data is also shown for reference.

The results in Table 1 shows that our method significantly outperform the baseline by 4.15 BLEU points on average. It also surpasses the system proposed in [Luong et al., 2015b] by 2.85 BLEU points. It’s also worth to mention that the improvement given by their method is lower than the reported one on the French to English translation task. A possible reason is that there are much more reorderings in Chinese English language pair, so it’s much harder to correctly predict which source word generate current target unknown word during translation. On the contrast, our model replaces rare words with similar words and keeps the completeness of the sentence, so that it is much easier for the translation model to learn correspondence between source and target words.

4.4 Comparison of Different Replacement Strategies
The performances of different replacement strategies are shown in Figure 3. It can be seen that considering bilingual context or bilingual similarity does improve the performance...
<table>
<thead>
<tr>
<th>System</th>
<th>03 (dev)</th>
<th>04</th>
<th>05</th>
<th>06</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses</td>
<td>28.68</td>
<td>29.87</td>
<td>27.27</td>
<td>29.17</td>
<td>28.77</td>
</tr>
<tr>
<td>Luong et al. (2015)</td>
<td>27.63</td>
<td>30.02</td>
<td>26.42</td>
<td>28.72</td>
<td>28.20</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>29.85</strong></td>
<td><strong>33.08</strong></td>
<td><strong>28.95</strong></td>
<td><strong>32.31</strong></td>
<td><strong>31.05</strong></td>
</tr>
</tbody>
</table>

Table 1: Translation results for different systems. Bahdanau et al. (2015) is the NMT model with attention mechanism, which is adopted as our baseline without replacement. Luong et al. (2015) is the approach to decorate target unk with alignment information.

Figure 3: Comparison of different replacement strategies. The performance is an averaged one over all NIST data sets we adopt.

no-rep: use original sentence without replacement
simple: use most similar word for replacement
bi-lm: use bilingual bi-directional language model to choose from top similar words
bi-sim: use bilingual word similarity to choose from top similar words

over the simple replacement method, although the magnitude is not so significant when compared with the improvement of the simple method over the baseline.

We also show the performance of translating original testing sentences without replacement in the figure. The result is quite impressive. Nearly 3 BLEU points can be achieved over the baseline if we use the NMT model trained on the bilingual data with rare words replaced, while keeping the testing sentence unchanged. The improvement is even larger than that brought by replacing the testing data. This demonstrates that training on complete sentences can greatly improve the quality of parameter estimation, and thus lead to much better translations.

4.5 Better Attention after Replacement

It is also interesting to check how replacement affect the translation process in detail. Figure 4 shows the translations for the same sentence by different systems. The figure on the left corresponds to the baseline model without replacement. Because the third word ‘呼吁’ (call) in the source sentence is outside the vocabulary, the baseline model cannot generate proper translation for it. What’s more, wrong attention to this rare word results in bad translation for the common word ‘美国’ (America). The baseline model add an extra word ‘north’ before ‘america’, which is not a translation of any word in the source sentence. And the last source word ‘对话’ (dialogue) is hardly attended by any target word, leading to missing translation for this word, although it is also a common word. On the contrary, our system find a similar word ‘呼吁’ to the source rare word, which is in fact a synonym to it. Given this complete sentence without any rare word, our system is able to generate a nearly perfect translation for the source sentence, in which all source words are properly attended and translated. Since the rare word ‘呼吁’ is not seen in the bilingual training corpus, the lexical translation table does not contain the translation for this word. So we keep the translation of the alternative word in the output. Last, the approach of [Luong et al., 2015b] generated a similar translation as the baseline system (not shown in the figure). And even if a target unk is generated and aligned correctly to the source rare word, the translation of the rare word still cannot be restored because it’s not in the lexical translation table.

4.6 Parameter Initialization

It is well known that parameter initialization has a big impact on the performance of neural networks. In this paper, we tried two ways to initialize the parameters of the system on replaced data. One is randomized initialization, the other is initializing with the parameter learned by the baseline model. According to our experience, the latter is robust and performs better than the former for our method. But for the approach proposed in [Luong et al., 2015b], the latter initialization strategy does not bring any benefit.

5 Analysis of Untackled Rare Words

Although our method can handle more than 90 percent rare words in the data, there are still some remain untackled, which can be divided into two categories as follows.

One is related with complex alignments. As described in section 3.1, we only handle one to one and one to zero (zero to one) mapping in this paper. there are also some one to many (many to one) and many to many alignments in the data. Here is an example:

```
还要标明引进批准文号

and indicating the import ratification number
```

the rare word ‘文号’ (document number) at the end of the source sentence aligns to two target word ‘ratification’ and ‘number’, and the target word ‘ratification’ also aligns to the second to last word ‘批准’ (ratification). If we focus
Figure 4: Better attention after replacement. Darker block denotes larger attention weight. Left: translation by baseline model; Right: translation by our model.

on word level replacement, then replacing both 'ratification' and 'number' with the translation of a word similar to '文号' will make the source word '批准' unaligned. So it’s better to do the replacement at phrase level. But how to find alternatives for phases remains a problem and it will be leaved as our future work.

The other class of untackled rare words are related with the similarity model. According to Zipf’s law [Zipf, 1949], it’s impossible to contain all words from a language in a corpus with limited size. And for speed and quality considerations, we also don’t train word vectors for words which appear less than 5 times in the mono-lingual data. So for those really rare words which are not seen or only seen a few times in the mono-lingual data, we cannot find words similar to them. According to our investigation, most of these really rare words belong to named entities, including number, person names, location names and organization names. With an extra named entity recognizer, we can replace these rare named entities with their type labels instead of similar words. And this will also be leaved as our feature work.

6 Related work

Neural machine translation has a short history of only a few years. [Kalchbrenner and Blunsom, 2013; Cho et al., 2014] first propose to use the encoder-decoder architecture to do sequence to sequence mapping. However, they only use it as an additional feature to evaluate the quality of phrase pairs in traditional machine translation. At the same time, [Sutskever et al., 2014] apply it in end-to-end machine translation. Having considered using only a single vector to represent source sentences with variable lengths is not reasonable, [Bahdanau et al., 2015] propose the attention mechanism to dynamically attend to different source words when generating different target words. [Luong et al., 2015a] propose to use local attention instead of global attention for improved speed and accuracy.

Different to traditional machine translation, NMT model can only employ a small vocabulary due to computational complexity. The rare words problem has attracted a lot of attention recently. Besides the work by [Luong et al., 2015b] which we compared in our paper, [Jean et al., 2015] propose to directly use large vocabulary with a method based on importance sampling. As pointed out in their paper, their method is complementary and can be used together with replacement methods.

In traditional machine translation, although all vocabulary in the training set can be used for decoding, there are still a lot of out-of-vocabulary words during testing and they hurt the translation performance a lot. Most work [Fung and Cheung, 2004; Marton et al., 2009; Jiang et al., 2007] addressing OOV problem focus on how to translate those OOV correctly during translation. They often resort to additional resources such as comparable data and synonym thesaurus. One notable exception is the work from [Zhang et al., 2012; 2013], which also focuses on the syntactic and semantic role of those OOV and propose to replace OOV with similar words during testing.

7 Conclusion

In this paper, we systematically studied how rare words impact the performance of NMT systems. And we proposed an effective approach of replacing rare words with similar in-vocabulary words. This approach not only enables the translation of rare words, but also reduces the ambiguity of the whole sentence, which is quite important for parameter estimation during training and in-vocabulary words translation during testing. Experiment results on Chinese to English translation tasks demonstrate the power of our methods. Our best replacement method outperforms the baseline by more than 4 BLEU points, which is also much better than the method proposed by previous work.

Acknowledgments

The research work has been partially funded by the Natural Science Foundation of China under Grant No. 61333018 and 91520204 as well, and it is also supported by the Strategic Priority Research Program of the CAS (Grant XDB02070007).
References


