Neural Name Translation Improves Neural Machine Translation

Xiaoqing Li[1,2] Jinghui Yan[4] and Jiajun Zhang[1,2] Chengqing Zong[1,2,3]

[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing,China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{xqli,jjzhang,cqzong}@nlpr.ia.ac.cn
[4] Beijing Jiaotong University, Beijing,China
17112083@bjtu.edu.cn

Abstract. In order to control computational complexity, neural machine translation (NMT) systems convert all rare words outside the vocabulary into a single unk symbol. Previous solution (Luong et al.[1]) resorts to use multiple numbered unks to learn the correspondence between source and target rare words. However, testing words unseen in the training corpus cannot be handled by this method. And it also suffers from the noisy word alignment. In this paper, we focus on a major type of rare words –named entity (NE), and propose to translate them with character level sequence to sequence model. The NE translation model is further used to derive high quality NE alignment in the bilingual training corpus. With the integration of NE translation and alignment modules, our NMT system is able to surpass the baseline system by 2.9 BLEU points on the Chinese to English task.

Keywords: Rare words · Named entity · Neural machine translation.

1 Introduction

Neural machine translation is a recently proposed approach to MT and has shown competing results to conventional translation methods (Kalchbrenner and Blunsom[2]; Cho et al.[3]; Sutskever et al.[4]). Despite several advantages over conventional methods, such as no domain knowledge requirement, better generalization to novel translations and less memory consumption, it has a significant weakness in handling rare words. In order to control computational complexity, NMT systems convert all rare words outside the vocabulary into a single unk symbol. Such conversion makes them unable to translate rare words. And those meaningless unks also increase the difficulty for the NMT model to learn the correspondence between source and target words.

To tackle this problem, Luong et al.[1] propose to augment the unk symbol with alignment information. Their method allow the NMT system to learn, for each

unk in the target sentence, the position of its corresponding word in the source sentence. A post-processing step is adopted to translate target unks with a dictionary.
This approach has been shown effective to handle rare words, but there are still some drawbacks. First, it cannot translate words outside the dictionary. Second, it relies on noisy word alignment. As known to all, automatic word alignment for rare words is far from perfect. Wrong alignment will reduce the quality of the bilingual corpus to train NMT model, and the dictionary extracted according to word alignment will also be noisy. Third, the content of rare words is totally ignored. Taking the following sentence as an example,

Judy chases the pat with unk .

the translation of this sentence into Chinese will be quite different depending on whether the last word is a person name or a modifier describing some feature of the pat.
To solve the above problems, we propose to translate rare words with character

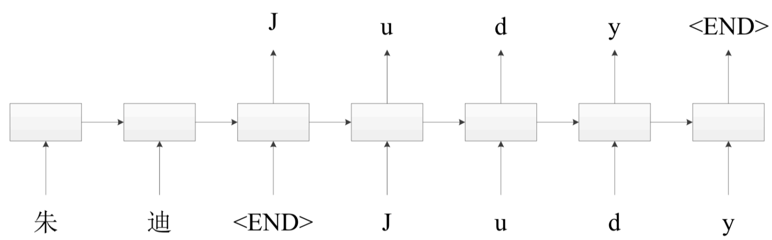

Fig. 1. Character level sequence to sequence model for NE translation.

level sequence to sequence model, as shown in Figure 1. Due to the limitation of existing resources, we limit our research in this paper to a major type of rare words–named entities. With an NE translator trained on external NE list, we can derive high quality NE alignment in the bilingual corpus, from which more NE pairs can be extract to further enhance the NE translator. Similar to Luong et al.[1], the aligned NE pairs are then replaced with their type symbols and an NMT model is trained on the new data. A post-processing step is employed to recover the translation of the replaced NEs.

Our experiments demonstrate the effectiveness of this approach. Evaluation on the Chinese to English translation task shows that our integrated system surpasses the baseline system by 2.9 BLEU points, and brings an improvement of 1.6 BLEU points over Luong's method.

2 Named Entity Translation and Alignment Model

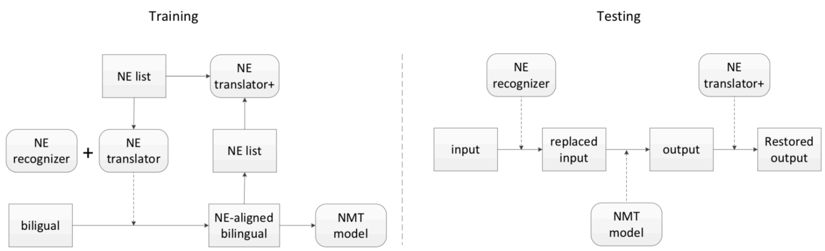

Fig. 2. System architecture to incorporate NE translation into neural MT.

Figure 2 gives an overview of the architecture of our system. In the training phase, we first train a neural NE translation system with a characterlevel sequence to sequence model. The initial training data consists of NE translation pairs, which can be obtained easily for many language pairs from the web. For example, we can extract linked Wikipedia titles and filter them according to category information. Then this NE translator together with a NE recognizer are used to find aligned NE pairs in the bilingual corpus. A list of NE pairs can be extracted from this corpus and it is combined with the original list to build a stronger NE translator. The aligned NE pairs in the bilingual corpus will be replaced with corresponding type symbols, resulting in sentence pairs like the following example,

ZH: LOC1 重新开放驻 LOC2 大使馆
EN: LOC1 reopens embassy in LOC2

Finally, this new data after replacement will be used to train an NMT model. In the testing phase, the NEs in the input sentence are first recognized and replaced with type symbols. After translated by the NMT model, the NE symbols in the output will be replaced with the translation of original NEs, which is generated by the NE translation module.

2.1 Named Entity Translator

The model we adopt to translate NE is a character level sequence to sequence model. It maps a source NE $s = (s_1, s_2, ..., s_m)$ into a target NE $t = (t_1, t_2, ..., t_n)$ with a single neural network as follows,

$$p(t|s) = \prod_{i-1}^{n} p(t_i|t_{<i}, s) \tag{1}$$

where the conditional probability is paramiterized with the encoder decoder framework. The encoder reads the source character sequence and encodes it into a sequence of hidden states. Then the decoder generates the target NE character by character based on the target hidden states. In this paper, we adopt the implementation of Bahdanau et al.[5] which introduces an attention mechanism while predicting each unit in target sequence. Please look into it for more detail. As pointed out by previous work, the total computational complexity will grow almost proportional to the target vocabulary in the sequence to sequence model. Fortunately, in the scenario of character level NE translation, the vocabulary size is only hundreds (such as English) to thousands (such as Chinese). While in the case of word level NMT, the vocabulary size is often hundreds of thousands to millions. So NMT systems usually limit vocab size to tens of thousands to make computation feasible. In character level model, there is no such need.

2.2 Named Entity Alignment

Named entity alignment based on bilingual corpus alone is a hard task. Since a lot of NEs appear only a few times in the corpus, we cannot collect enough statistical evidence to infer reliable alignments with traditional word alignment model. Previous work (Huang et al.[6]; Feng et al.[7]) design multiple features, including translation score, transliteration score, distortion score etc., and use iterative training to discover aligned NEs in bilingual corpus.
However, if we have a high quality and high coverage NE list, together with a powerful NE translation model, things could be much easier. We can learn a NE translator from the list, then use it to translate recognized NEs in one language and compare it with word sequence (up to trigram) in the other language. The longest common subsequence between NE translation candidate and target word sequence is adopted for similarity matching in this paper,

$$LCS(c, s) = \frac{1}{2}(|c| + |t| - ED(c, t)) \tag{2}$$

$$sim(c, t) = LCS(c, t)/|c| \tag{3}$$

where ED is the edit distance and |x| is the length of x. For example, the longest common subsequence between 'bolin' and 'berlin' is 'blin'.
The final NE alignment result is the union of bi-directional matching, i.e. matching source NE with target word sequence and vice versa. We do not match bilingual NEs directly because the automatic NE recognition is not good enough and some NEs are not recognized in each language.
It has to mention that we don't have a list of numerical and temporal expressions to train a corresponding translation model before alignment. To calculate the similarity score, we carry the following conversion for both Chinese and English numerical expression,

– Convert Chinese and English numbers one to nine to abric numbers, discard all other characters. For example, '百分之四点二' (4.2%) will be converted into 42, and 4,200 will also be converted into 42.
– Add a few rules to handle exceptions such as month.

The above conversion is used for the sake of alignment. After NE pairs are extracted according to the alignment, they are used to train the NE translation model to handle NEs in testing data.

3 Experiments

We evaluate our method on the Chinese to English translation task. Translation quality is measured by the BLEU metric (Papineni et al., 2002[8]).

3.1 Settings

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 sentence pairs longer than 50 words either on the source side or on the target side are discarded. The initial NE pairs are extracted from the Wikipedia, which consist about 350k entries. We use an in-house developed NE recognizer for Chinese and Stanford NER (Finkel et al.[9]) for English.
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.

3.2 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. Names inside the vocabulary are not handled. The number of hidden units is 1,000 for both the encoder and decoder. And the embedding dimension is 500 for all source and target tokens. The network parameters are updated with the adadelta algorithm and the learning rate is set to 10^{-4}. The above setting is used both in the character level NE translation and word level sentence translation.

3.3 Name Translation and Alignment Performance

Because the recognition performance and crosslingual consistency are not good for organization names, we ignore this type and only handle numerical/temporal expressions, person names and location names in this paper. To evaluate the NE translation performance, we randomly extract 100 instances from the NIST testing data for each type, and manually find their translations in the reference. Whereas the NE alignment performance is evaluated on the same amount of

Table 1. Translation and alignment performance with neural NE translation model

	N/T	PER	LOC
translation	0.71	0.28	0.35
trans. + lex. table	0.78	0.48	0.70
alignment	0.97	0.93	0.96

samples extracted from the training data. The results are shown in Table 1.
It could be seen from the table that the translation accuracy is relatively low. Since accuracy is calculated in word level, the NE translation is regarded wrong even when there is only one letter different. In order to improve the NE translation accuracy in the post-processing step, we propose to use the lexical table extracted from the bilingual data if the NE could be found, otherwise the neural NE translation model will be used. On the other hand, the alignment accuracy is quite high. And most alignment errors relate to wrong word segmentation or NE recognition according to our investigation.

3.4 Sentence Translation Performance

Table 2. Translation results for different systems

System	03 (dev)	04	05	06	Average
baseline	25.65	28.94	25.13	27.86	26.90
unk rep.	27.63	30.02	26.42	28.72	28.20
NE rep.	27.90	30.67	28.20	29.42	29.05
unk+NE rep.	**29.01**	**31.33**	**28.80**	**30.08**	**29.80**

We compare the translation performance of our method with that of Luong et al.[1]. The results are shown in Table 2. The baseline system we adopt is the attention-based model proposed in Bahdanau et al.[5] . It can be seen that only replacing rare NEs with our method results in a better performance than replacing all rare words with Luong's method. After combining the two methods, i.e., replacing NEs with our method and replacing other rare words with Luong's method, we could obtain an extra performance boost of 0.75 BLEU points, and the final performance surpass the baseline by 2.9 BLEU points on average. It has to be mentioned that Luong's method is not as effective on the Chinese-English language pair as reported on the French-English language pair. A possible reason is the automatic word alignment quality is worse on the former language pair.

4 Related Work

Inability to handle rare words is a significant defect of NMT systems. And it has attracted much attention recently. Besides the work of Luong[1], Jean et al.[10] 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. Sennrich et al.[11] propose to represent rare words as sequences of subword units, and compares different techniques to segment words into subwords. Wang et al.[12] use a hierarchical structure for NMT, in which word representations are derived from character representations.

The problem of NE translation has been studied for a long time. Knight and Graehl[13] study it with probabilistic finite-state transducers. Li et al.[14] present a joint source-channel model for direct orthographical mapping without intermediate phonemic representation. Freitag and Khadivi[15] propose a technique which combines conventional MT methods with a single layer perceptron. Deselaers et al.[16] use deep belief networks for machine transliteration. There are also some previous works trying to integrate NE translation into traditional MT systems. Hermjakob et al.[17] study the problem of "when to transliterate". Li et al.[18] propose to combine two copies of training data, the original one and the one with aligned NEs replaced.

5 Conclusion

In this paper, we enhance the ability of NMT system to handle rare words by incorporating NE translation and alignment modules. With the help of extra NE list and NE recognizer, our method is able to produce high quality NE alignments, and thus improves the data quality to train NE translation and sentence translation model. Experimental results show that our approach can significantly improve the translation performance.

References

1. Luong, M.T., Sutskever, I., Le, Q.V., Vinyals, O., Zaremba, W.: Addressing the rare word problem in neural machine translation. arXiv preprint arXiv:1410.8206 (2014)
2. Kalchbrenner, N., Blunsom, P.: Recurrent continuous translation models. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. (2013) 1700-1709
3. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
4. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. (2014) 3104-3112
5. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

6. Huang, F., Vogel, S., Waibel, A.: Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In: Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition (2003) 9-16
7. Feng, D., Lv, Y., Zhou, M.: A new approach for English-Chinese named entity alignment. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. (2004) 372-379
8. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics (2002) 311-318
9. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd annual meeting on association for computational linguistics, Association for Computational Linguistics (2005) 363-370
10. Jean, S., Cho, K., Memisevic, R., Bengio, Y.: On using very large target vocabulary for neural machine translation. arXiv preprint arXiv:1412.2007 (2014)
11. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909 (2015)
12. Ling, W., Trancoso, I., Dyer, C., Black, A.W.: Character-based neural machine translation. arXiv preprint arXiv:1511.04586 (2015)
13. Knight, K., Graehl, J.: Machine transliteration. Computational linguistics **24**(4) (1998) 599-612
14. Li, H., Zhang, M., Su, J.:A joint source-channel model for machine transliteration. In: Proceedings of the 42nd Annual Meeting on association for Computational Linguistics (2004) 159-166
15. Freitag, D., Khadivi, S.: A sequence alignment model based on the averaged perceptron. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (2007) 238-247
16. Deselaers, T., Hasan, S., Bender, O., Ney, H.: A deep learning approach to machine transliteration. In: Proceedings of the Fourth Workshop on Statistical Machine Translation, Association for Computational Linguistics (2009) 233-241
17. Hermjakob, U., Knight, K., Daumé III, H.: Name translation in statistical machine translation-learning when to transliterate. Proceedings of ACL-08: HLT (2008) 389-397
18. Li, H., Zheng, J., Ji, H., Li, Q., Wang, W.: Name-aware machine translation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (2013) 604-614