The Impact of Named Entity Translation for Neural Machine Translation

Jinghui Yan[1], Jiajun Zhang[2,3], and Jinan Xu[1] Chengqing Zong[2,3,4]

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

Abstract. Named entity translation has been shown in many studies that could have positive impact on performance of sentence level neural machine translation. In this paper, we study a mainstream structure that incorporating an external named entity translation model to neural machine translation. We give several comparison experiments by applying different named entity translation model structures, to clearly represent the impact of this structure in improving quality of neural machine translation. The experiments show that the proposed approach is able to achieve posistive result on some datasets and we give our analysis of influence factors.

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

1 Introduction

Neural machine translation (NMT) achieved impressive result in recent years. Thanks to a series creative proposals of [1–3], such as "sequence-to-sequence" and "attention mechanism", there is a notable improvement in aspects of sentence accuracy and fluency. However, existence of out-of-vocabulary (OOV) is still a problem that neural machine translation always suffers from. For reducing computation complexity, the number of vocabulary of NMT system has to restricted to a limited size, due to which lots of rare words have to be replaced by *unk* symbols. Named entity, always playing as low-frequency words, is the main cause of OOV problem.

Many studies have been proposed to handle this problem, and one iconic work is [4], which presents a subword-level neural machine translation model based on the byte pair encoding (BPE) algorithm[5]. BPE is a compression algorithm that originally developed for word segmentation, it can restrict vocabulary into a fixed size by segment word to subword, for instance, a rare named entity "Estelle" would be transfer into 'Es tell e'. Though BPE could really transfer a rare

word to a sequence of frequent subwords and generate a subword vocab file for a training corpus, there are still enormous amount of NE that we can not capture. Wang et al.[6] discusses the advantages and disadvantages of different translation granularities in Chinese-English NMT, but it does not lays emphasis on which granularity is the most suitable for named entity. Especially for transliteration words of named entities, there are innumerable combinations of subwords that a finite training set could never be covered totally.

To tackle the problems above, we propose to translate the named entities prior to the translation of whole sentence by an external named entity translation model. Li et al.[7] followed the "tag-replace" training method proposed by Loung et al.[8], using character level sequence to sequence model to translate named entities and an NMT model is trained on the new data of which named entities have been replaced with their type tags. In this paper, we propose to use different named entity translation models and a more general named entity alignment method to test the feasibility of this scheme. To verify the effectiveness of the scheme, we manually force all of named entities be exactly translated and the result will be present in Section 4.

2 Related work

The research of named entity translation has come a long way. Earlier researchers focus on constructing a grapheme mapping table from source named entity A to a corresponding character of target B. Wan et al.[9] made mainly two mapping steps to translate English country names into Chinese names: English phoneme to Chinese pinyin, and Chinese pinyin to han characters. Some researchers try to use statistical methods to directly learn syllable alignment probabilities from bilingual named entity corpus. Li et al.[10] transliterated person names from English to Chinese used modified source-channel model for direct orthographic mapping to generated probabilistic rules from a bilingual dictionary. Asif et al.[11] using modified source-channel model that incorporates different linguistic knowledge of possible conjuncts for Bengali and English. Yang et al.[12] presented a two-step CRF model for machine transliteration. Recently, deep learning achieve impressive results for sentence level machine translation, inspired by which, researchers began to use neural machine translation model to automatically catch the features for NE translation. Li et al. [7] split both Chinese and English words into character level, using a "sequence-to-sequence" structure for named entity translation. Li et al.[13] segment the English words into subword level, building a subwords-to-characters model for English-Chinese person name transliteration and get an impressive result.

For the study of incorperating named entity tranlsation into neural machine tranlstion system, the mainstream approach is making NEs translation as a separeting procedure from the original NMT model. Li et al.[7] trained a sentence level NMT model with NEs in parallel sentences tagged by NE symbols, then the NE symbols in output sentences will be replaced with corresponding translations, which is generated by its enternal NE translation module. Wang et al.[14]

proposed similar method to Li[7] but only focus on PER named entities and only use the extracted parallel person names from training data. Their result brings no significant improvement but they claim it will be useful for human evaluation.

3 System Scheme

To verify the impact of named entity translating on neural machine translation, we propose to separate the NMT translating processing into two parts: named entity translating and other parts of sentences translating. The overview of our architecture shows in Figure 3.

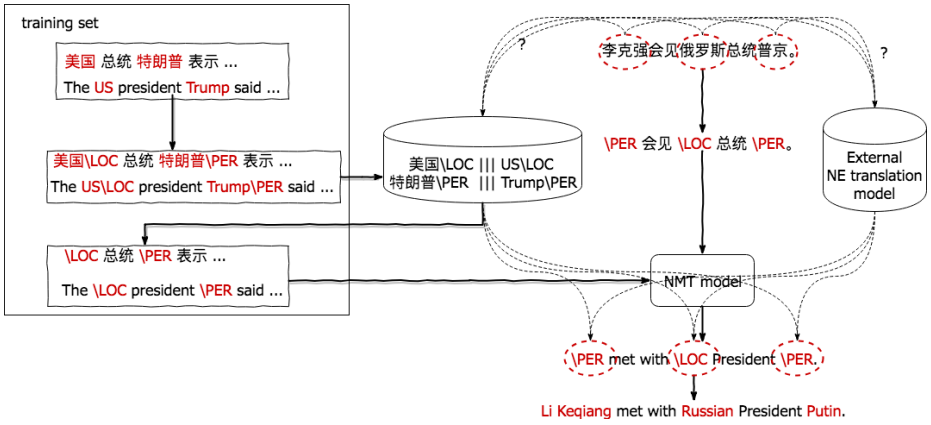

Fig. 1. System architecture of incorporating NE translation into neural MT

Tag Replacing Prior to the training process of NMT, a word aligner together with NE recognizer are used to replace the named entities which parallelly appear in the training corpus with special tags as follows:

- "LOC1 总统 PER1 表示..."
- "The LOC1 president PER1 said ..."

We use two monolingual NE recognizer on both sides of training set to recognize all NE inside corpus. Considering a named entity may contain multi-gram tokens, we need recombine them into unigrams. For instance, using underline to recombine "青沙公路" and "Tsing Sha Highway" to "青沙 _ 公路/LOC" and "Tsing_Sha_Highway/LOC" respectively. Then, a word alignment tool can be used to align and extract those pairs with same NE tags.

As shown in Figure 3, a bilingual parallel NE dictionary can be generated by above steps, which is used to replace the NE appearing in parallel sentences pairs with tags.

External NE Translation In order to adapt named entity translation to sequence-to-sequence (seq2seq) translation model, we need cut the word into more granular patterns. Here we build two kind of Chinese-English named entity seq2seq translation models. The first one is a character-to-character model which we split both Chinese and English parts into sequence of characters. For another one, we use BPE algorithm to segment named entities in English side into sequence of subwords. Then a character-to-subword NE translation model can be built. Figure 1 and Figure 2 show the architectures of our two models respectively.

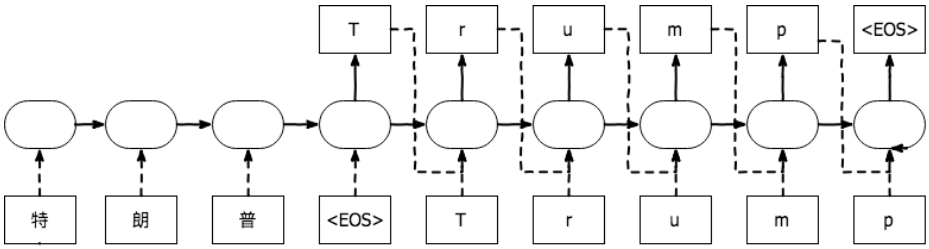

Fig. 2. Character-to-character NE translation model

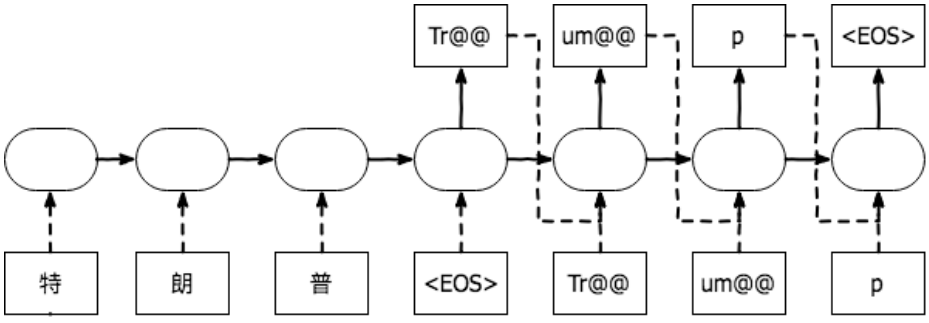

Fig. 3. Character-to-subword NE translation model

4 Experiment

4.1 Data Set

We use CWMT Chinese-English parallel data to train our NMT model, which contains about 9M sentences pairs. We choose NIST03 dataset as development and NIST 04-06 are used as test set. Meanwhile, LDC Chinese-English NE translation corpus[17] (LDC2005T34) which is compiled from Xinhua News Agency newswire texts, is used for building our named entity translation model. The entire LDC2005T34 corpus contains several categories of NEs (e.g., person, location, organization, press, industry). In our experiment, we only focus on person name, location and organization. After extracting these three categories of NEs, we split the train, dev, test set by their proportions Table 1 shows the statistics.

Table 1. Statistics of the data for building external NE translation model

Category	Number of Entries			
	Total	Train	Dev	Test
PER	693,705	692,265	720	720
LOC	230,844	230,364	240	240
ORG	37,409	37,329	40	40
Total	961,958	959,958	1,000	1,000

4.2 Training Detials

We limit both the baseline NMT system and the tagged NMT system with vocabulary size of 30k Chinese subwords and 35k English subwords. We use Transformer[15] implementation of a self-attention based sequence-to-sequence NMT model. The batch size is 4096, and drop-out probablity is 0.3. We use Adam gradient optimization[18] with learning rate 0.1. We stop training at 200k steps for both models when accuray of tunning sets are not improved for both models. For our named entity translation system, we use same model structure with baseline NMT model. We split all Chinese named entities into characters. In character-to-character NE translation system, we split English named entities into characters too, except that we use '@' to replace the original blank inside those entities. For character-to-subword NE translation system, English NE are segmented into subwords by using BPE model, and we set the size of subword into 35k.

4.3 Performance and Results

NE Translation System We totally bulid five NE translation models to test their performence in translating different kinds of named entities. As shown in

Table 1, we use total training data to built the character-to-character model (c2c-Total) and character-to-subword model. Also, we train three character-to-character models (c2c-PER, c2c-LOC, c2c-ORG) using each of the three categories individually to verify if a clearly classified named entity training model could be helpful. We only use accuracy to evaluate the result. Table 2 shows the evaluation result.

Table 2. Performence of different NE translation model

	PER	LOC	ORG
c2c-Total	37.5%	29.8%	29.2%
c2sub-Total	34.2%	27.3%	19.5%
c2c-PER	37.5%	-	-
c2c-LOC	-	28.2%	-
c2c-ORG	-	-	19.5%

Due to the mulit-transliteration problem, all of five models give a poor performance of accuracy, lower than 40%. Compared vertically, character-to-character model performs a little bit better than character-to-subword model. We believe it is because that the subword segmentation can not completely include all combinations of English character sequences. In other words, the vocabulary of target side is still limited. Also, we can see that, building the model by category individually will not help performance much, but somhow drop a bit. We think there are two reasons here:
First, the NE data of different categories always nest inside each other. For example, the "Harvard University" is an organization named entity, however, part of the phrase "Harvard" could also be nested inside the person named entity "John Harvard" which would provide helpful translation information when translating "Harvard University". Second, there are lots of transliteration words in all kind of categories of named entities. For instance, the LOC named entity "Cassandra" and the PER named entity "Sandra" are share a common transliteration part "sandra". Therefore, it may be an unwise choice to individually build translation model by the category.
By horizontally comparing, it could be found that translation performence of PER named entity always performs better than others. Chen[16] make analyses for Chinese-English named entity corpus LDC 2005T34, and she found that the transliterated entities take up 100 % of PER, transliterated location names account for 89.4 % of all LOC, and transliterated organization names take only 12.6 % of ORG named entites. Accroding to Chen's statistics, we surmise that better performence of PER entities are mainly due to a high rate of transliteration words, of which rules are easier to be learned.
More specifically, we check the named entities that are erroneously translated by tag-replace system. Table 3 shows some false examples. The first raw gives a typical example of wrong PER translating—the wrong domain. Obviously it is

possible to translate 布希 into “Bouchy” if only considering the pronunciation, however, in some areas like Taiwan, people use 布希 a set of phrases to refer to president Bush. Raw 2 and raw 3 show another unavoidable problem—the NE translation model can not cover all rules of combinations of named entities, especially for the non-transliteration named entities. Because of the absent of combination “中国-China” or “南韩-south korean” inside the ex-NE dict, the NE translation model transliteration it character-for-character by using “中-naka; 国-kuni” and “南-nan; 韩-ham”.

Table 3. Example of erroneously translated NE by tag-replace system

	NET	baseline
布希	Bouchy	Bush
中国	Nakakuni	China
南韩	Nan ham	south korean

Named Entity Alignment We use phrase-based SMT system Moses[5] to train named entity alignment. As the method mentioned in Section 3.1, we first recognize all named entities inside both source and target training sets, then we recombine them as a unigram token and put a tag followed. All aligned pairs are extracted from the phrase-table generated by Moses. We make a constraint of both source-target and target-source unigram phrase translating probability higher than 0.3. We totally extract 18,620 aligned pairs of named entities and Table 4 gives the statistic and its performence. Raw 3 and raw 4 are numbers of named entities recognised by Chinese and English monolingual named entity recognizer respectivly.

Table 4. Performence of Named entity alignment

	PER	LOC	ORG
Recognised (zh)	149,437	111,877	147,471
Recognised (en)	170,861	121,294	327,011
Extracted (pairs)	8,924	4,159	5,537
Accuracy	98%	94%	94%

To evaluate the alignment performence, we randomly select 100 extracted pairs of each NE category and analyse them manually. It could been seen from Table 4 that the alignment accuracy of all categories are pretty well. We also count the numbers of those extracted NE pairs tagged in the training sets.

[5] http://www.statmt.org/moses

Sentence Translation Performance Accroding to the result in Section 4.3, we use only c2c-Total model in our external NE translation (ex-NET) system. The total Chinese-English named entity corpus LDC 2005T34 is used as our external named entity dictionary (ex-NE dict) and we use ex-NET system only when its corresponding translation can not be found in ex-NE dict. We train two NMT models, one of which we replace all three kinds of named entity (PER, LOC, ORG) inside training data with tags, the other of which we recongize PER named entities only. Table 5 shows the results of above two models in the second and third raw respectively.

Table 5. Performence of NMT system (BLEU score)

	03 (dev)	04	05	06	Avg.
baseline	40.33	40.53	40.43	39.66	40.23
tag-replace	35.33	37.98	37.91	35.67	36.72
tag-replace_PER_only	39.88	**41.53**	**40.61**	**40.83**	**40.71**
UNK	31.43	33.81	33.13	29.93	32.07

Table 6. Frequency of NE appeared in dev and test sets

	03	04	05	06
NE	1,874	2,983	2,219	1674
lines	919	1,788	1,082	1,000
words	23,534	49,151	29,355	23,917

Table 7. Comparision of NE translating performance in dev set.

	BLEU		Accuracy		
			PER	LOC	ORG
baseline	40.33		32%	64%	47%
tag-replace	UNK	31.43	0%	0%	0%
	Oracle	40.58	100%	100%	100%
tag-replace_PER_only	UNK	36.71	0%	64%	47%
	Oracle	40.83	100%	64%	47%

Unfortunately, all test sets give depressing results to tag-replace system. However, in the third raw of the table, things are different. Except the development set, all other three test sets present improvements of BLEU score. For test sets 04 and 06, each of the increasing score are more than 1 BLEU scores, which can be seen as significant improvement.

More detaily, we count the frequency of NE appeared in each set in Table 6, which gives a intuitive feeling that there are numerosity named entities in test sets. To find the extent of the impact of the named entity translation for NMT, we build two control groups for both "tag-replace" and "tag-replace_PER_only" systems using development data set. We first replace all recognized named entities in translated output sentences with "UNK" tags, which means, for "tag-replace" system, all of three categories of named entities are not be translated. And for "tag-replace_PER_only" system, only PER named entities are not be translated and others keep the same translation accuracy with baseline system. The "UNK" group could be seen as the lower bound of the system. To find the upper bound of system, we manually replace all named entities with their correct translations, which we call the "Oracle" group. The result shows in Table 7 — the accuracy of named entity translation of each group is calculated by comparing with the "Oracle" group. Comparing the two control groups, it prove that, named entity translation could deeply impact the quality of NMT. However, when comparing the "Oracle" group with the baseline, a notable phenomenon is that there is no expected improvement of "Oracle" group even we manually set 100% accuracy of named entity translation. Moreover, we keep the "Oracle" group of "tag-replace_PER_only" system the same accuracy of "LOC" and "ORG" with baseline and 100% accuracy of "PER", which perform better BLEU scores than the "Oracle" group of "tag-replace" system, though the latter has both higher accuracy of "LOC" and "ORG". Therefore, the unsatisfactory performence of "tag-replace" can not be merely attributed to the bad quality of external NE translation system

Sentences Translation Prior to decoding procedure, same NE recognizer would be used to extract all named entities in input sentences and replace them with corresponding tags. Then, the system searches the NE dictionary generated by sextion 3.1 to find if there is given translations, otherwise the external NE translation model would be used to give the corresponding translations. As Figure 3 shows, each of the tags in the NMT output will be directly replaced by its NE translating.

5 Conclusion

In this paper, we propose to verify the feasibility of separating the neural machine translation processing into two parts: named entity translation, and the other parts of sentences translating. The experiment proves that the quality of named entity translation will affect the final performence of whole sentences translating and the quality of named entity translation largely depend on if there are same domain knowledge between training set and the input namd entities. The "tag-replace_PER_only" system in which we only recognize PER named entities inside sentences get a postive result of BLEU scores in neural machine translation However, the "tag-replace" system which handle all categories of named entities

seems less effective in using named entity translation to improve quality of NMT system. Since the oracle system does not get a significant improvement, the unsatisfactory performence of "tag-replace" can not be merely attributed to the bad quality of external NE translation system. There could be a structural defects of "tag-replace" and we plan to find other influential factors of the system in the future.

References

1. Kalchbrenner, N., Blunsom, P.: Recurrent continuous translation models. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. (2013) 1700-1709
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
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. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909 (2015)
5. Gage, P.: A new algorithm for data compression. The C Users Journal **12**(2) (1994) 23-38
6. Wang, Y., Zhou, L., Zhang, J., Zong, C.: Word, subword or character? An empirical study of granularity in Chinese-English NMT. In: China Workshop on Machine Translation, Springer (2017) 30-42
7. Li, X., Zhang, J., Zong, C.: Neural name translation improves neural machine translation. arXiv preprint arXiv:1607.01856 (2016)
8. 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)
9. Wan, S., Verspoor, C.M.: Automatic English-Chinese name transliteration for development of multilingual resources. In: Proceedings of the 17th international conference on Computational linguistics-Volume 2, Association for Computational Linguistics (1998) 1352-1356
10. 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, Association for Computational Linguistics (2004) 159-166
11. Ekbal, A., Naskar, S.K., Bandyopadhyay, S.: A modified joint source-channel model for transliteration. In: Proceedings of the COLING/ACL on Main conference poster sessions, Association for Computational Linguistics (2006) 191-198
12. Yang, D., Dixon, P., Pan, Y.C., Oonishi, T., Nakamura, M., Furui, S.: Combining a two-step conditional random field model and a joint source channel model for machine transliteration. In: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration, Association for Computational Linguistics (2009) 72-75
13. Li, Z., Chng, E.S., Li, H.: Named entity transliteration with sequence-to-sequence neural network. In: Proceedings of the Asian Language Processing (IALP), 2017 International Conference on, IEEE (2017) 374-378
14. Wang, Y., Cheng, S., Jiang, L., Yang, J., Chen, W., Li, M., Shi, L., Wang, Y., Yang, H.: Sogou neural machine translation systems for WMT17. In: Proceedings of the Second Conference on Machine Translation. (2017) 410-415

15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. (2017) 5998-6008
16. Chen, Y., Zong, C., Su, K., et al.: Joint Chinese-English named entity recognition and alignment. Chinese Journal of Computers **34**(9) (2011) 1688-1696
17. Huang, S.: Ldc2005t34: Chinese<-> English named entity lists v 1.0. Linguistics Data Consortium (2005)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)