Incorporating External Annotation to improve Named Entity Translation in NMT

Maciej Modrzejewski Thanh-Le Ha Alexander Waibel
Institute for Anthropomatics and Robotics
KIT - Karlsruhe Institute of Technology, Germany
maciej.modrzejewski@student.kit.edu
firstname.lastname@kit.edu

Miriam Exel Bianka Buschbeck
SAP SE, Walldorf, Germany
firstname.lastname@sap.com

Abstract

The correct translation of named entities (NEs) still poses a challenge for conventional neural machine translation (NMT) systems. This study explores methods incorporating named entity recognition (NER) into NMT with the aim to improve named entity translation. It proposes an annotation method that integrates named entities and inside–outside–beginning (IOB) tagging into the neural network input with the use of source factors. Our experiments on English→German and English→ Chinese show that just by including different NE classes and IOB tagging, we can increase the BLEU score by around 1 point using the standard test set from WMT2019 and achieve up to 12% increase in NE translation rates over a strong baseline.

1 Introduction

The translation of named entities (NE) is challenging because new phrases appear on a daily basis and many named entities are domain specific, not to be found in bilingual dictionaries. Improving named entity translation is important to translation systems and cross-language information retrieval applications (Jiang et al., 2007). Conventional neural machine translation (NMT) systems are expected to translate NEs by learning complex linguistic aspects and ambiguous terms from the training corpus only. When faced with named entities, they are found to be occasionally distorting location, organization or person names and even sometimes ignoring low-frequency proper names altogether (Koehn and Knowles, 2017).

This paper explores methods incorporating named entity recognition (NER) into NMT with the aim to improve NE translation. NER systems are often adopted as an early annotation step in many Natural Language Processing (NLP) pipelines for applications such as question answering and information retrieval. This work explores an annotation method that integrates named entities and inside–outside–beginning (IOB) (Ramshaw and Marcus, 1999) tagging into the neural network input with the use of source factors. In our experiments, we focus on three NE classes: organization, location and person, and use the state-of-the-art encoder-decoder Transformer network. We also investigate how the granularity of NE class labels influences NE translation quality and conclude that specific labels contribute to the NE translation improvement. Further, we execute an extensive evaluation of the MT output assessing the influence of our annotation method on NE translation. Our experiments on English→German and English→Chinese show that by just including different NE classes and IOB tagging, we can increase the BLEU score by around 1 point using the standard test set from WMT2019 and achieve up to 12% increase in NE translation rates over a strong baseline.

2 Related Work

Several research groups propose translating named entities prior to the translation of the whole sentence by an external named entity translation model. Li et al., (2018a); Yan et al., (2018); Wang et al., (2017) follow the "tag-replace" training method using an external character-level

En	BPE only	Belfast - Gi@@ ants won thanks to Patri@@ ck D@@ w@@ yer
En	fine-grained	Belfast_2 -_0 Gi@@_3 ants_3 won_0 thanks_0 to_0 Patri@@_1 ck_1 D@@_1 w@@_1 yer_1
En	coarse-grained	Belfast_1 -_0 Gi@@_1 ants_1 won_0 thanks_0 to_0 Patri@@_1 ck_1 D@@_1 w@@_1 yer_1
En	IOB tagging	Belfast_B -_O Gi@@_B ants_I won_O thanks_O to_O Patri@@_B ck_I D@@_I w@@_I yer_I
En	Inline Ann. (fine-grained)	<LOC> Belfast </LOC> - <ORG> Gi@@ ants </ORG> won thanks to <PER> Patri@@ ck D@@ w@@ yer </PER>

Table 1: Different annotation configurations; i. fine-grained: (0) for a regular *sub-word* (default), (1) for NE class *Person*, (2) for NE class *Location*, (3) for NE class *Organization* ii. coarse-grained: (0) default, (1) to denote a NE

sequence-to-sequence model to translate named entities. Li et al. (2018b) explore inserting inline annotations into the data providing information about named entity features. Such annotations are inserted into the source sentence in form of XML tags, consisting of XML boundary tags and NE class labels.

Recently, researchers have shown the benefit of explicitly encoding linguistic features, in form of source factors, into NMT (Sennrich and Haddow, 2016; García-Martínez et al., 2016). Dinu et al. (2019) use source factors successfully to enforce terminology. The work of Ugawa et al. (2018) is similar to ours, in the way that they also incorporate NE tags with the use of source factors into the NMT model to improve named entity translation. They, however, introduce a chunk-level long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) layer over a word-level LSTM layer into the encoder to better handle compound named entities. Furthermore, they use a different network architecture (LSTM), and apply a different annotation technique (IO tagging) than we explore (IOB tagging). Finally, the work at hand provides an extensive evaluation of NE quality translation (Section 5.2), including a human assessment (Section 5.3).

3 NMT with NE tagging

We explore incorporating NE information as additional parallel streams (source factors) to signal NE occurrence in the fashion described in Sennrich and Haddow (2016). Source factors provide additional word-level information, are applied to the source language only, and take form of supplementary embeddings that are either added or concatenated to the word embeddings. This is illustrated with the following formula:

$$E \cdot x = \bigoplus_{f \in F} E_f \cdot x_{if} \quad (1)$$

where $\bigoplus \in \{\sum, \|\}$, $(\cdot)$ denotes a matrix-vector multiplication, E_f is a feature embedding matrix, x_i is the i-th word from the source sentence, and F is a finite, arbitrary set of word features. While we use a state-of-the-art encoder-decoder Transformer network, our approach does not modify the standard NMT model architecture, thus can be applied to any sequence-to-sequence NMT model.

Further, we also explore whether the NE class granularity may influence translation quality and help decrease word ambiguity. For this purpose, we define a "fine-grained" case, where we use specific NE class labels (e.g. person, location, organization) and also a "coarse-grained" case, where we use two different source factor values only: (0) as default and (1) to denote a named entity in a generic manner. Additionally, we investigate whether inside–outside–beginning (IOB) tagging (Ramshaw and Marcus, 1999) used to signalize where a NE begins and ends as a second input feature may guide models to translate compound named entities better. In IOB tagging, (B) indicates the beginning, (I) the inside and (O) the outside of a NE (a regular word or a sequence of words).

We annotate source sentences with an external NER system. Examples for the different annotation strategies (that we experiment with) are presented in Table 1. Each sub-word is assigned an index denoting its corresponding source factor value.

As our goal resembles that of Li et al. (2018b), we compare our approach against their inline annotation method with XML boundary tags. Li et al. (2018b) use specific NE class labels, which correspond to the "fine-grained" case in our work. We refer to their approach as "Inline Ann. (fine-grained)" and present this annotation method in Table 1.

4 Experiments

4.1 Parallel data & pre-processing

We train NMT systems for English→German and English→Chinese on data of the WMT2019 news

	En→De	En→Zh
No. of sentences	2,146,644	2,128,234
No. of sentences with NE	1,082,873	1,153,545
Percentage	≈ 50.44%	≈ 53.95%
ORG labels	983,558 (53%)	1,325,462 (57%)
PER labels	223,309 (12%)	211,892 (9%)
LOC labels	639,304 (35%)	796,269 (34%)

Table 2: Occurrences of NE annotations in the training datasets

Label type	Variant	IOB	En→De	En→Zh
fine-grained	sum	no	**33.61**	26.29
fine-grained	concat 8	yes	33.11	**26.45**
fine-grained	sum	yes	33.07	26.26
coarse-grained	concat 8	yes	32.90	26.08
coarse-grained	sum	yes	32.70	26.34
Baseline		no	32.60	26.29
Inline Ann. (fine-grained)		no	32.50	26.05

Table 3: BLEU scores on *newstest2019* (WMT2019)

translation task.[1] For English→German we use the data from Europarl v9 and news commentary data v14. For English→Chinese the models are trained on news commentary v14 and UN Parallel Corpus v1.0. The latter dataset is shortened to match the size of the training dataset for English→German by using the newest data from the end of the corpus for training, see also Table 2.

As NE Recognition is an active research field and the search for best recognition methods continues, the quality of NER systems may vary under different research scenarios and domains (Goyal et al., 2018). Incorrect NE annotation in the data may influence the results of this work negatively. Therefore, we focus on three well-researched NE classes: *Person*, *Location* and *Organization*, limiting, thus, the possibility of incorrect annotation.

We use spaCy Named Entity Recognition (NER) system[2] to recognize named entities in the source sentences. The ratio of sentences in the training data with at least one named entity occurrence (based on three NE classes) in the source sentence amounts to 50.44% for En–De and 53.95% for En–Zh. Table 2 presents the details.

We tokenize the English and German corpora using the spaCy Tokenizer[3], and use the OpenNMT Tokenizer[4] (mode aggressive) on the Chinese side. Further, we perform a joint source and target Byte-Pair encoding (BPE) (Sennrich et al., 2016) for English→German and disjoint for English→Chinese, both with 32,000 merge operations. For every source sentence in the training data (after applying BPE), we generate two files with source factors: i. marking named entities (either the coarse-grained or the fine-grained case), ii. marking IOB tagging. The baseline model is trained with no external annotation.

[1]http://www.statmt.org/wmt19/translation-task.html
[2]https://spacy.io/usage/linguistic-features\#named-entities
[3]https://spaCy.io/api/tokenizer
[4]https://github.com/OpenNMT/Tokenizer

4.2 NMT architecture

We use the Sockeye machine translation framework (Hieber et al., 2017) for our experiments and train our models with a Transformer network (Base) (Vaswani et al., 2017) with 6 encoding and 6 decoding layers all with 2048 hidden units. We use word embeddings of size 512, dropout probability for multi-head attention of size 0.1, batch size of 4096 tokens, a maximum sequence length of 100 and source factor embedding of size 8 for the concatenation case. Each model is trained on 1 GPU Tesla T4. Training finishes if there is no improvement for 32 consecutive checkpoints on the validation data *newstest2018* (validation data from the WMT2019 news translation task).

5 Results

5.1 General translation quality

We perform the evaluation on the standard test dataset *newstest2019* from the WMT2019 news translation task. It has identical content for En–De and En–Zh and contains 1997 sentences, in which 63.95% of the sentences on the English side contain at least one named entity. There are 2681 named entity occurrences; 908 belong to the label *Location* (34% of all NEs), 870 to the label *Person* (32%) and 903 to the label *Organization* (34%); annotated with spaCy NER. Each sentence with named entity occurrence contains, on average, approx. 2 NEs. To assess the general translation quality, we calculate the BLEU score using the evaluation script *multi-bleu-detok.perl* from Moses (Koehn et al., 2007). We detokenize the MT output with *detokenizer.perl* (Koehn et al., 2007) for En–De and use OpenNMT *detokenize* function to do the same for En–Zh.

Table 3 displays the results. Column "Label type" denotes whether specific ("fine-grained") or generic ("coarse-grained") NE labels are used; column "Variant" describes whether source factors are added ("sum") or concatenated ("concat") to

En→De						
Label type	Variant	IOB	LOC	PER	ORG	Total
fine-grained	sum	no	73.68	70.11	61.79	69.89
fine-grained	concat 8	yes	72.87	**71.96**	63.41	70.67
fine-grained	sum	yes	**75.71**	70.85	**69.11**	**72.39**
coarse-grained	concat 8	yes	74.09	71.22	62.60	70.67
coarse-grained	sum	yes	75.30	71.22	65.04	71.61
Baseline		no	74.09	71.59	60.16	70.36
Inline Ann. (fine-grained)		no	70.45	67.16	61.79	67.39

Table 4: Results of the automatic in-depth analysis on *random300* dataset for En–De with spaCy NER, *NE match rate* in %

En→Zh						
Label type	Variant	IOB	LOC	PER	ORG	Total
fine-grained	sum	no	**41.67**	20.07	31.62	24.41
fine-grained	concat 8	yes	33.33	**23.36**	36.76	**27.96**
fine-grained	sum	yes	**41.67**	20.44	33.09	25.12
coarse-grained	concat 8	yes	33.33	22.63	33.09	26.30
coarse-grained	sum	yes	33.33	21.90	**38.97**	27.73
Baseline		no	33.33	18.98	35.29	24.64
Inline Ann. (fine-grained)		no	33.33	19.71	34.56	24.88

Table 5: Results of the automatic in-depth analysis on *random300* dataset for for En–Zh with Stanford NER, *NE match rate* in %

the word embeddings; column "IOB" describes whether IOB tagging is used as a second source factor stream.

Almost all models annotated with source factors show improvements w.r.t BLEU in comparison to the baseline; with one En–Zh model being insignificantly worse. Overall, the fine-grained model with source factors added and no use of IOB tagging seems to perform best and achieves around one BLEU point more than the baseline (for En–De). As the BLEU score only assesses the quality of NE translation indirectly, we do not deem it to be a reliable evaluation metric to assess the NE translation quality. As named entities affect only a small part of a sentence, we do not expect high BLEU variations and continue with the in-depth named entity analysis in the next section.

5.2 Automatic hit/miss NE evaluation

In this section we execute an automatic in-depth analysis of NE translation quality with spaCy (German models) and Stanford NER (Finkel et al., 2005) (Chinese models). For this purpose, we randomly select 100 sentences from *newstest2019* containing at least one named entity for each of the three classes (PER, LOC, ORG) on the English side of the corpus, in total 300 sentences. We refer to this dataset in later part of this work as *random300*. We annotate the reference sentence with an external NER system (spaCy or Stanford NER) to find named entities and compare if they appear in the hypothesis in the same form (string-based). If yes, we define this case as a "hit", otherwise as a "miss" and calculate the result according to the *NE match rate* formula: $\frac{hit}{hit+miss}$. Table 4 and Table 5 display the results. Column "Total" calculates the accumulated *NE match rate* for three named entity classes.

At first glance, we see that the result values for En–De are significantly higher than for En–Zh. We attribute this to the transliteration issues which emerge while translating from English to Chinese and, thus, occurring mismatch between the reference and hypothesis translation. In general, the baseline models show high performance as a certain amount of NEs has already been seen by the network in the training data. Furthermore, we observe improvements in named entity translation for En–De and En–Zh among almost all classes, showing that augmenting source sentences with NE information leads to their improved translation. There is, however, no consistent improvement in the models not using IOB tagging annotation. Their total *NE match rate* values are lower than that one of the baseline models. As such, IOB tagging, indicating compound named entities, proves to be an important piece of information for the NMT systems. Further, augmenting the model with exact NE class labels (fine-grained case) seems to achieve higher *NE match rates* in comparison to the coarse-grained case. Additionally, coarse-grained models perform better than the baseline. This finding indicates that the mere information that a word is a NE proves to be useful to the NMT system even if the class is not clearly specified. Inline Annotation does not deliver promising results, contrary to the findings of Li et al. (2018b), with the total *NE match rate* below that one of the baseline system (En–De) or insignificantly above (En–Zh).

Validation of the *NE match rates* After having executed the automatic in-depth analysis with spaCy NER, we wish to validate the results of the En–De models with a second state-of-the-art NER system: Stanford NER. The analysis is conducted in an identical way as earlier and only the En–De models are analyzed. At the point of writing this paper, spaCy does not provide a Chinese model. Table 6 presents the results. Column "Total" cal-

En→De						
Label type	Variant	IOB	LOC	PER	ORG	Total
fine-grained	sum	no	76.25	76.14	60.00	73.70
fine-grained	concat 8	yes	75.62	77.16	64.62	74.88
fine-grained	sum	yes	**80.00**	**78.68**	**69.23**	**76.78**
coarse-grained	concat 8	yes	75.62	77.66	67.69	75.36
coarse-grained	sum	yes	77.50	76.65	**69.23**	76.48
Baseline		no	78.75	76.65	60.00	74.64
Inline Ann. (fine-grained)		no	73.75	74.11	60.00	71.80

Table 6: Results of the automatic in-depth analysis on *random300* dataset for En–De with Stanford NER, *NE match rate* in %

culates the accumulated *NE match rate* for three named entity classes.

First, we observe that the overall *NE match rates* are higher than in Table 4. We attribute this phenomenon to the fact that Stanford NER recognizes a different set of NEs in the reference sentences than spaCy does. This, however, is not problematic as we are interested in the variations in *NE match rates* between the models. In general, there are no differences in the results of the automatic in-depth analysis, regardless whether spaCy or Stanford is used to conduct it. All models trained with IOB tags translate NEs more accurately than the baseline model does. Again, fine-grained model trained with IOB tags and source factors added to the word embeddings achieves the highest *NE match rate*. The model trained without IOB tags has a lower *NE match rate* than the baseline re-confirming thus the usefulness of the IOB tags.

5.3 Human hit/miss NE evaluation

As NER systems are prone to delivering inaccurate results,[5] we also perform a human evaluation. It consists in recognizing NEs in the reference translation, comparing them to the corresponding NE translation in the MT output and calculating the *NE match rate* on the *random300* dataset. We compare the baseline and the best model (highest total *NE match rate* in Tables 4 and 5) for En–De and En–Zh and refer to them as *annotated* models. If a NE is in a different form in the hypothesis than the reference proposes or a NE is transliterated into or from Chinese, but its form is still grammatically and semantically correct, its occurrence is counted as correct. Human evaluation is executed by one native speaker for each language pair. Table 7

[5]spaCy's German model has 83% F1-Score (https://spaCy.io/models/de) with a warning that it may "perform inconsistently on many genres", the same holds for Stanford NER: https://nlp.stanford.edu/projects/project-ner.shtml.

En→De						
Label type	Variant	IOB	LOC	PER	ORG	Total
fine-grained	sum	yes	93.02	83.52	78.01	85.17
Baseline		no	89.77	82.05	70.92	82.14
En→Zh						
fine-grained	concat 8	yes	73.85	67.04	64.27	68.05
Baseline		no	71.43	61.90	57.35	63.24

Table 7: Results of the human in-depth evaluation on *random300* dataset, *NE match rate* in %

presents the results of the human hit/miss evaluation. Column "Total" calculates the accumulated *NE match rate* for three named entity classes.

The *NE match rate* for human hit/miss evaluation is higher than for its automatic counterpart. This is due to the fact that all false positives in the reference and false negatives in the hypothesis are eliminated. Most importantly, we can state that the *annotated* models perform consistently better than the baseline and, in fact, the incorporation of external annotation in form of source factors into the source sentence leads to an improvement in NE translation. There is an increase of 3.67% in the total *NE match rate* value for En–De and 7.61% for En–Zh. Furthermore, we observe the greatest *NE match rate* improvement when translating organizations' names (+9.99% for En–De, and +12.07% for En–Zh).

5.4 Accuracy of spaCy NER

While executing the human hit/miss NE evaluation, we also annotated false positives and false negatives in the reference, executing, thus, a quality check of spaCy NER on data from the news domain (on *random300* dataset, German model only). Precision value is 84.43% and recall amounts to 85.93%. The above observation leads to the conclusion that incorrect NE annotation may occur relatively frequently in the training data. We hypothesize that NE annotation with source factors may lead to better results if the training data is fully correctly annotated.

5.5 Discussion

In this section we discuss our observations based on the human evaluation and provide translation examples. The use of source factors seems to alleviate the problem of ignoring low-frequency proper names as the *annotated* models appear to consistently react to NE occurrence by producing a translation. The baseline, however, may ignore more complex NEs, producing, thus, under-

Source	Palin, 29, of Wasilla, Alaska, was arrested (...) according to a report released Saturday by **Alaska State Troopers**.
Reference	Palin, 29, aus Wasilla, Alaska, wurde (...) verhaftet. Gegen ihn liegt bereits ein Bericht (...), so eine Meldung, die am Samstag von den **Alaska State Troopers** veröffentlicht wurde.
Annotated	Palin, 29 von Wasilla, Alaska, wurde (...) verhaftet (...), wie ein am Samstag von **Alaska State Troopers** veröffentlichter Bericht besagt.
Baseline	Laut einem Bericht von **Alaska**, der Samstag veröffentlicht wurde, wurde Palin, 29 von Wasilla, Alaska, (...) verhaftet (...).
Source	Saipov, 30, allegedly used a **Home Depot** rental truck (...).
Reference	Saipov, 30, hat (...) angeblich einen Leihwagen von **Home Depot** (...) benutzt (...).
Annotated	Saipov, 30, soll einen Mietwagen aus dem **Home Depot** benutzt haben (...).
Baseline	Saipov, 30, soll einen **Home Department Depot** Rental benutzt haben (...).
Source	The pair's business had been likened to **Gwyneth Paltrow's Goop** brand.
Reference	Das Geschäft der beiden war mit der Marke **Goop** von **Gwyneth Paltrow** verglichen worden.
Annotated	Das Geschäft des Paares wurde mit der Marke **Gop** von **Gwyneth Paltrow** verglichen.
Baseline	Das Geschäft des Paares wurde mit der Marke von **Gwyneth Palop** verglichen.
Source	The **Giants** got an early two-goal lead through strikes from Patrick Dwyer and **Francis** Beauvillier.
Reference	Die **Giants** hatten durch Treffer von Patrick Dwyer und **Francis** Beauvillier eine frühe Zwei-Tore-Führung.
Annotated	Die **Giganten** bekamen durch die Streiks von Patrick Dwyer und **Franziskus** Beauvillier ein frühes Ziel.
Baseline	Die **Giganten** erhielten durch die Streiks von Patrick Dwyer und **Francis** Beauvillier ein frühes Ziel.

Table 8: Translation examples: Comparison of the *annotated* model and baseline for En–De

translation as in the *Alaska State Troopers* example in Table 8. Furthermore, source factors seem to guide the *annotated* models better (in comparison to the baseline) to prevent over-translation, as shown in the *Home Depot* example or miss-translation (*Gwyneth Paltrow's Goop*), both examples are in Table 8.

On the other hand, a frequent cause of errors in the *annotated* models stems from the fact that organizations' or persons' names are translated verbatim instead of being kept in their original forms, as in the *Francis/Franziskus* and *Giants/Giganten* example in Table 8. This problem concerns both the *annotated* model and the baseline. This behavior may not be desirable for persons' names, yet for organizations' names the desired output is dependent on the context and translation language pairs.

6 Conclusion

Our work focused on establishing if annotating named entities with the use of source factors leads to their more accurate translation. We can state that the general translation quality with the *annotated* models improves (improvements in BLEU score). Additionally, in-depth automatic and human named entity evaluation prove that the same holds true for NE translation.

The accuracy of named entity annotation plays a crucial role during the annotation of named entities in the training data as well as during evaluation (automatic hit/miss analysis). By establishing spaCy's F1-Score on *random300* during the human hit/miss analysis to amount to approx. 85%, we conclude that the accuracy of any NER system greatly influences the practicability of our approach. Therefore, the improvement of named entity translation is closely related to the improvement of NER systems.

Acknowledgements

We would like to thank Zihan Chen for her help with the human evaluation of the En–Zh translation.

References

Dinu, Georgiana, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

García-Martínez, Mercedes, Loïc Barrault, and Fethi Bougares. 2016. Factored neural machine translation. *arXiv preprint arXiv:1609.04621*.

Goyal, Archana, Vishal Gupta, and Manish Kumar. 2018. Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29:21–43.

Hieber, Felix, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. LSTM can solve hard long time lag problems. In *Advances in neural information processing systems*, pages 473–479.

Jiang, Long, Ming Zhou, Lee-Feng Chien, and Cheng Niu. 2007. Named entity translation with web mining and transliteration. In *Proceedings of the 20th international joint conference on Artifical Intelligence*, pages 1629–1634.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the Association for Computational Linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Li, Xiaoqing, Jinghui Yan, Jiajun Zhang, and Chengqing Zong. 2018a. Neural name translation improves neural machine translation. In *China Workshop on Machine Translation*, pages 93–100. Springer.

Li, Zhongwei, Xuancong Wang, Aiti Aw, Eng Siong Chng, and Haizhou Li. 2018b. Named-entity tagging and domain adaptation for better customized translation. In *Proceedings of the Seventh Named Entities Workshop*, pages 41–46.

Ramshaw, Lance A and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Sennrich, Rico and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Ugawa, Arata, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural machine translation incorporating named entity. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3240–3250.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, Yuguang, Shanbo Cheng, Liyang Jiang, Jiajun Yang, Wei Chen, Muze Li, Lin Shi, Yanfeng Wang, and Hongtao Yang. 2017. Sogou neural machine translation systems for wmt17. In *Proceedings of the Second Conference on Machine Translation*, pages 410–415.

Yan, Jinghui, Jiajun Zhang, JinAn Xu, and Chengqing Zong. 2018. The impact of named entity translation for neural machine translation. In *China Workshop on Machine Translation*, pages 63–73. Springer.