Fast-Paced Improvements to Named Entity Handling for Neural Machine Translation

Pedro Mota, Vera Cabarrão, Eduardo Farah
Unbabel, Lisbon, Portugal
{pedro.mota, vera.cabarrao, eduardo.farah}@unbabel.com

Abstract

In this work, we propose a Named Entity (NE) handling approach to improve translation quality within an existing Natural Language Processing (NLP) pipeline without modifying the Neural Machine Translation (NMT) component. Our approach seeks to enable fast delivery of such improvements and alleviate user experience problems related to NE distortion. We implement separate NE recognition and translation steps. Then, a combination of standard entity masking technique and a novel *semantic equivalent* placeholder guarantees that both NE translation is respected and the best overall quality is obtained from NMT. The experiments show that translation quality improves in 38.6% of the test cases when compared to a version of the NLP pipeline with less-developed NE handling capability.

1 Introduction

NE play a crucial role in many downstream NLP tasks. There is extensive research showing that properly handling NE improves the performance of systems performing Question Answering (Talmor and Berant, 2019), Summarization (Zhou et al., 2021), and Information Retrieval (Wang et al., 2021). In this paper, we focus on NMT, another task that benefits from NE modeling (Shavarani and Sarkar, 2021). NMT models are prone to disturb NE, leading to critical quality issues in the translation. Overcoming such problems is challenging since it is hard to have good coverage of all possible entities in the training data. This is due to the open-ended nature of NE as well as their domain specificity. For example, for the `Organization (ORG)` category, new entities appear daily in a variety of domains. Moreover, NE are linguistically complex structures that can occur in ambiguous contexts. This impairs the ability of models to generalize and instead learn unwanted biases (Hassan Awadalla et al., 2018; Modrzejewski et al., 2020). This causes NE to be hallucinated towards frequent realizations, omitted, or incorrectly translated. Figure 1 shows some examples of this issue in the output translation of an English $\rightarrow$ French NMT model. This occurs despite the model having 65×10^6 parameters and being trained with 100 million sentence pairs.

The NMT community has long been familiar with the NE handling problem (Koehn and Knowles, 2017). This has spurred research on how to address such model limitations. Invariably, all works resort to either incorporating new modeling features in existing NMT architectures (Li et al., 2019; Modrzejewski et al., 2020) or integrating with external knowledge sources to bridge the NE gap (Zhao et al., 2020a; Feng et al., 2021).

In spite of the achievements of the previously mentioned works, they have the drawback of requiring a model-specific solution. In a commercial setting, this is problematic since NE handling, at least for some categories, might come only as an afterthought. Having the NLP pipeline already in place entails that rolling out changes can be slow due to the high number of existing models. It should be noted that there are also time and budget constraints regarding the model size and volume of training data in order to make a NMT system economically viable. This blocks translation quality improvements related to NE handling.

	DATE distortion
Input:	However, on *18 February 2022* you again contacted us.
Translation:	Cependant, le *18 février*, vous nous avez à nouveau contactés.
	PERSON distortion
Input:	Hi *Zéphyrin*
Translation:	Bonjour *Zécerin*

Figure 1: Examples of NE distortions by NMT.

In this paper, we propose an alternative perspective to NE handling. We argue that it is important to deliver, as fast as possible, translation quality improvements to end-users, avoiding critical communication issues. To achieve this, we describe a process that enables NE handling to be deployed in an NLP pipeline without changing the NMT component. In an NMT industry scenario, this is relevant since flexibility in model architecture is necessary to accommodate different use cases. Thus, the decoupling of NE handling is desirable to not add extra requirements to the NMT component.

In particular, we first carry out a NE recognition pre-processing step. Then, we obtain the corresponding translation for that entity. Finally, we resort to a semantically-equivalent mask that the NMT can properly handle. When it is not possible to generate a semantically-equivalent entity, we default to the standard placeholding method from NMT. This affords a good trade-off between translation quality and the NLP pipeline run-time.

2 Related Work

The standard approach to NE handling within a NLP pipeline corresponds to introducing NE information and forwarding it to the NMT component. The end goal is to allow the model to improve the NE translation quality. In previous work, there are different approaches to make use of this NE information, which we summarize below.

A possible approach is the placeholder method (Wang et al., 2017; Post et al., 2019), where source sentences are masked by a generic entity token, exposed to the NMT model during training. After translation, the masks are placed back into the target sentence, based on an index or alignment. Li et al. (2016; 2019) extend this approach to overcome the limitation of dealing with rare words in this setting. This is done with a dedicated character-level sequence-to-sequence model for NE translation. A NE recognition step is also added to enable the use of category-specific entity tokens. NEs are crawled from the training data and their translation extracted from Wikipedia. The NE translation pairs are then used to train both the character-level and NMT models.

Another line of research uses entity embeddings to convey word-level NE category information to guide the NMT model. An example is source factors (Sennrich and Haddow, 2016), which take the form of supplementary embeddings that are added or concatenated to existing word embeddings in the model. Ugawa (2018) combines this with an additional Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1996) layer to better handle NE. This contrasts with the work from Modrzejewski et al. (2020), where better translation quality is achieved by directly combining source factors in a Transformer network (Vaswani et al., 2017). SemKGE (Moussallem et al., 2019) take a similar approach but construct the embeddings differently. These map subject-relation-object triples from a Knowledge Graph (KG) (Vrandečić and Krötzsch, 2014) into a continuous vector space to obtain Knowledge Graph Embeddings (KGE) (Bordes et al., 2013). To this end, a supervised fastText (Joulin et al., 2017) classifier determines a set of referring expressions of NE from the KG and uses them to initialize the embedding weights of the NMT model. Zhao et al. (2020b) use a similar methodology but focus on dealing with the drawback of only taking into account NE that appear in both the KG and the training dataset. To leverage the remaining relevant information in the KG, phrase translation pairs are first extracted from the training data. The pairs that appear in the KGs are considered seed pairs in a KGEs semantic space. This semantic space is then used to compare new NE with the seed pairs. If KGE are close, then a synthetic sentence pair is generated by replacing the original NE with the new ones.

Continuing the entity embedding research line, Xie et al. (2022) take it a step further and provide a generic recipe to achieve a single end-to-end NE-aware NMT model, which avoids the overhead of separate NE handling steps. Moreover, there is no extra cost at inference time since the NE components can be disabled. To achieve this, an enhanced encoder and decoder are trained in a multi-task framework by combining translation and NE recognition in a focal loss (Lin et al., 2017).

Given the current state-of-the-art, we conclude that previous approaches introduce coupling to the NMT architecture by either changing it or jointly training new embeddings. While this brings advantages in many scenarios, we argue that it is also valuable to address the use case where a large NLP pipeline already exists and fast incremental improvements to NE need to be delivered by means of new categories. In this context, we build upon the placeholder approach, where we are willing to sacrifice *translation quality* for a *translation guarantee* that certain words are perfectly translated. We extend this approach to better reconcile these two competing aspects as well as study the more complex case where the NE require translation.

3 Named Entity Handling

In our approach, we first start by performing a NE recognition pre-processing step (Section 3.1). Then, we obtain the corresponding translation for a given target language (Section 3.2). We forward all the previous information to a NE handler step that obtains the best possible quality from the existing NMT model while guaranteeing that the expected translation appears in the output translation.

3.1 Recognition

For this step, we combined regex and neural network-based approaches to identify NE in a source sentence. This way we can capture NE with a structured format as well as context dependent ones. We support the following categories:

- **Regex**: `GLOSSARY`, `IP-ADDRESS`, `EMAIL`, `ALPHANUMERIC-ID`, `PHONE-NUMBER`, `BANK-NUMBER`, `CURRENCY`, `NUMBER`, `PERCENTAGE`, `URL`, and `DATE` (numerical).

- **Neural**: `PERSON`, `COUNTRY`, `Products & Organizations (PRO-ORG)`, and `DATE` (alphanumerical).

The `GLOSSARY` category is a manually curated list of terminology that must be enforced in a particular domain. The `ALPHANUMERIC-ID` captures NEs such as promotional codes. The regex `DATE` category matches numerical dates (e.g.: `yy/mm/dd`). The neural `DATE` covers the numbers and text case such as "January 1^{st}, 2022". The `PRO-ORG` is a merge between two different categories, Products, and Organizations since it is often the case that they are almost indistinguishable.[1] The remaining categories are self-explanatory.

3.2 Translation

Different translation needs stem from the different possible NE categories as well as the language pair. For a set of categories, the NE should be kept as in the original text and should not be translated. This is the case for `URL`, `PRO-ORG`, `PHONE-NUMBER`, `IP-ADDRESS`, `BANK-NUMBER`, (generally) `PERSON`, `NUMBER`, `PERCENTAGE`, `DATE` (numerical), `CURRENCY`, `ALPHANUMERIC-ID`, and `EMAIL`.

When the NE cannot be copied, it is necessary to provide a suitable translation. For cases where the NE can be translated without context, a dictionary-based approach can be suitable. This is the case for the `COUNTRY` category since there is a limited number of possible realizations. Moreover, building the dictionaries for a variety of language pairs is feasible through available resources such as KGs. Another option can be to outsource the NE translation to an external NMT provider such as Google, Amazon, or Microsoft. A use case for this is the `DATE` (alphanumerical) category since there is some variety in the day, month, and year structure as well as language-specific punctuation rules that make it hard to translate. Using an external service can be a solution in this case, because the provider can afford to have very large generic models (trained on large amounts of data), making them more robust to some NE categories.

Depending on the target language, a category might require or not translation. This is the case of `PERSON`, which requires transliteration if the source and target scripts do not match, namely in Arabic, Russian, and Greek. The strategies described above can still be applied. The dictionary approach can be supported by character transliteration tools when a name cannot be found.

[1] For example, the search engine *Google* is also the name of the organization.

3.3 Neural Machine Translation Integration

The output of the previous steps is a set of word spans with the NE category and expected translation. In the next section, we describe how to integrate this output with NMT to obtain a more robust NE handling strategy.

3.3.1 Named Entity Masking

It is plausible that a particular realization of a NE will not be present in the training data of the NMT model, leading to a poor quality translation. For example, the `PERSON` category has a wide variety of realizations since it varies according to the language, can have abbreviations, and many possible combinations of first, second, and last names.

To overcome the previous problem, we propose the use of a *semantic equivalent* version to mask the original NE. This is akin to the standard masking in NMT, which corresponds to a context-free replacement of a class of input tokens with a single mask token. The idea is to collapse distributionally similar tokens into a single token that the decoder can then be trained to reliably copy to the translation. Then, a demasking step replaces the token placeholder with either a copy of the source match value or the translation obtained from a dictionary. This feature is commonly available in NMT industry to satisfy the requirement of being able to enforce domain-specific terminology. The advantage of using a semantic equivalent mask is that it does not change the underlying meaning of the sentence. Thus, we can avoid degrading the translation quality in other parts of the sentence since the NMT has access to all the necessary linguistic information. To achieve this we only need to search for a semantic equivalent that the NMT is likely to correctly translate. To this end, we came up with a list of plausible candidates and empirically observe if NMT was able to translate them.

Despite increased translation robustness, there is still no guarantee that the NMT will output the semantic equivalent mask. When this is the case, we argue that it is likely that NMT distorted the mask. To repair the translation we trigger an entity fallback mechanism. This mechanism resorts to standard masking using the available default entity token placeholder. This is also useful in situations where generating a semantic equivalent is not possible. For example, for the `COUNTRY` category, one can easily find the necessary translation for a variety of languages. The obstacle is that gender is hard to obtain, especially because it depends on the target language. Thus, we can first check if the raw sentence translation contains the expected NE translation; if it does not, then resort to entity fallback. The drawback of this strategy is that it will hide linguistic information from the NMT. Thus, errors such as agreement in gender are expected.

The previously described strategy achieves improvements on both translation *quality* and translation *guarantee* aspects. This occurs because we use a semantic equivalent mask to have the best possible quality from the existing NMT and only resort to the entity fallback guarantee after checking that the expected translation was not output.

3.3.2 Semantic Equivalent Generation

To apply the previous strategy, it is necessary to define a semantic equivalent NE generation process. This is not straightforward since the required linguistic features might not be available and vary across categories and languages. For example, for the `PERSON` category, we need to determine the gender (female, male, or unisex). Despite being an open-ended NE, it is still possible to get good coverage by leveraging resources available online.[2] From these resources, we can build a name lookup table with the gender information. For `PERSON` NE containing more than one word, we heuristically check each word in the lookup table and return the first match. Another linguistic feature that the `PERSON` category can have is if it corresponds to a family name. Although we do not try to identify this feature, we generate a semantic equivalent family name if we find a title (e.g.: “Mr.”; “Mrs.”).

Putting all NE handling steps together, we provide two examples of our approach in Figure 2. In the `PERSON` category example, the semantic equivalent masking was able to repair the NE distortion described in the beginning of this paper (Figure 1). In the `COUNTRY` category example, the NMT did not output the expected translation, causing a critical error. After re-translating with the default entity token `$MASK`, we were able to guarantee that “Japão” appeared in the final output. It should be noted that there is an agreement error: the preposition “na” is in the feminine form and it should be in the masculine one (“no”). Despite this error, this is less critical than omitting the NE, and, thus, the overall translation quality was improved.

[2]For example, `https://github.com/lead-ratings/gender-guesser`

PERSON Example	
Input:	Hi Zéphyrin
NE Recognition:	Hi **Zéphyrin**
NE Translation:	Hi **[Zéphyrin→Zéphyrin]**
Semantic Equivalent:	Hi **[Thomas→Thomas]**
NMT:	Bonjour **Thomas**
Output:	Bonjour **Zéphyrin**
COUNTRY Example	
Input:	I understand that currently you are in Japan
NE Recognition:	I understand that currently you are in **Japan**
NE Translation:	I understand that currently you are in **[Japan→Japão]**
NMT:	Entendo que, atualmente, está no país
Retranslation:	I understand that currently you are in **[$MASK→Japão]**
	Entendo que, atualmente, está na **$MASK**
Output:	Entendo que, atualmente, está na **Japão**

Figure 2: NE handling pipeline.

4 Experiments

We carry experiments in all NE handling steps, namely: recognition (Section 4.1), NE translation (Section 4.2), and NMT integration (Section 4.3).

4.1 Named Entity Recognition Experiments

The following sections describe the evaluation of NE recognition step.

4.1.1 Experimental Setup

Our NLP pipeline is deployed in a commercial setting, thus, there are requirements constraining the model to have a small memory footprint and fast inference time. The architecture of the neural network is a stack combining GloVe word embeddings (Pennington et al., 2014), an LSTM layer, a hierarchical character-level BiLSTM-CRF (Lample et al., 2016), and a final CRF (Lafferty et al., 2001) layer on top. We use word embeddings of size 100 and the remaining layers have 256 hidden units. Training runs for up to 120 epochs, on batch size 32, and learning rate 0.1.

The training data is from the customer support domain, in the travel, technology, and education topics. The data was annotated by a linguist expert, taking approximately 3 weeks. In total, 46168 English sentences were annotated. This experiment focuses on the following categories: PERSON, COUNTRY, PRO−ORG, and DATE. The number of instances for each categories is: 5968, 397, 695, 17057, and 2178, respectively.

We compare our performance with two out-of-the-box models: spaCy 3.2.1 (Honnibal et al., 2020), *en_core_web_sm* model, and Stanza 1.3.0 (Qi et al., 2020), OntoNotes-based model. To measure performance, we use precision, recall, and F_1 metrics in a 10-fold cross-validation setup.

4.1.2 Experimental Results

The results are depicted in Table 1 and show that our custom model performs significantly better than the out-of-the-box models, with differences up to 72.6 in F_1. Between spaCy and Stanza, we observe that the latter generally performs better. It is also possible to observe that there are some NE categories that are easier to recognize for our custom model. This is the case of PERSON, and DATE, which shows that there is a lot of structure for these categories in our domain. In the remaining categories, the main issues we detected were variance in context (PRO−ORG), making it hard for the model to generalize, and a low number of occurrences (COUNTRY, and DATE), limiting the ability to learn the category during training.

Given the previous results, we conclude that in our use case of customer support domain it is worth paying the acquisition cost of the manually annotated NE data since it provides a great performance boost over out-of-the-box models.

4.2 Named Entity Translation Experiments

We now report the experimental results for the NE translation step.

Category	Metric	spaCy	Stanza	Custom
PERSON	Pre	35.7±1.9	71.1±3.4	97.4±0.8
	Rec	57.1±2.4	56.8±1.4	97.4±2.8
	F_1	43.9±1.9	63.1±1.8	96.3±1.6
COUNTRY	Pre	23.4±5.7	61.9±5.3	93.1±4.0
	Rec	7.5±1.9	6.2±2.5	76.5±8.9
	F_1	11.2±2.4	11.1±4.2	83.7±5.9
PRO-ORG	Pre	40.8±2.9	62.4±3.8	85.9±1.6
	Rec	30.8±1.3	36.2±1.5	88.4±2.5
	F_1	35.0±1.2	45.8±1.8	87.1±1.4
DATE	Pre	25.4±2.3	31.4±2.8	87.7±9.1
	Rec	78.6±4.5	63.7±2.7	95.3±2.3
	F_1	38.4±2.8	41.9±2.6	91.0±5.1

Table 1: NE recognition experimental results.

4.2.1 Experimental Setup

As mentioned in Section 3.2, in language pairs with different scripts, like English → Russian, the PERSON category might need translation. In this context, we collected 784 sentences containing the PERSON category and asked a Russian native speaker to provide the transliteration. Then, we measured the accuracy performance for the following approaches: one-to-one character mapping, Polyglot (Chen and Skiena, 2016), name dictionary (Merhav and Ash, 2018), and NMT providers (Google, Amazon, and Microsoft). In the name dictionary approach, we fallback to character mapping if the name is not in the dictionary.

4.2.2 Experimental Results

The results in Table 2 show that the most competitive approaches are the name dictionary and Google, with an accuracy up to 31.9% higher. For the name dictionary approach, we observe that the majority of the errors occur (95.3%) when the name was not present in the dictionary, resulting to a fall back to the character mapping strategy.

	%Accuracy
Character Mapping	50.4
Polyglot	46.2
Name Dictionary	**82.3**
Google	81.3
Amazon	75.8
Microsoft	74.5

Table 2: Name translation results.

The main difficulty we observed in this task stems from the fact that name transliteration needs to follow very specific rules. These introduce many exceptions to the standard character mapping, which explains its low results. An example of such rules is that the character "ы" can never go at the end of a name ("й" should be used instead). This makes the standard mapping from "y" fail for names like "Rey".

4.3 Neural Machine Translation Experiments

To understand the impact on quality of extending our NLP pipeline with new categories, we performed several experiments for the PERSON, COUNTRY and DATE (alphanumerical) categories.

4.3.1 Experimental Setup

The datasets are from the same domain as in previous experiments and the evaluations were done by expert linguists with fluent knowledge of the language pairs evaluated. To this end, we marked if the translation was better, the same, or worse than the previous version of the pipeline. We consider that the quality is better if errors in the original NMT are corrected or if the translation is more adequate. We consider translations as the same if both are equivalent. Finally, we consider that translations are worse if new errors are introduced. The experiments were carried out in a total of 2130 sentences in 7 language pairs (English source).

Regarding the baseline NMT, we trained bilingual models following the training procedure for the Transformer-base architecture (Vaswani et al., 2017). We first train a generic model using data available in the Opus platform (Tiedemann, 2012); the data volume is in the order of magnitude of hundreds of millions. Then, the model is fine-tuned with domain data; the data volume is in the order of magnitude of hundreds of thousands. The improved NE handling used the semantic equivalent (PERSON), Google NMT provider (DATE), and dictionary (COUNTRY) translation strategies.

4.3.2 Experimental Results

The obtained results are described in Table 3. Overall, it can be observed that the percentage of improved sentences is higher than the percentage of damaged sentences across all categories and languages. This validates that our NE handling strategy is beneficial. The majority of the cases marked as worse are due to incorrectly identified NE in the recognition step.

Category	Target	%Better	%Same	%Worse	#Sentences
PERSON	German	14.9	80.1	4.9	141
	French	22.5	77.4	0.0	31
	Dutch	45.4	38.3	16.1	99
	Brazilian	93.3	5.15	1.52	330
DATE	German	59.5	25.6	14.8	168
	French	68.5	21.1	10.3	194
	Portuguese	65.0	20.4	14.5	240
COUNTRY	German	8.3	87.9	3.8	346
	French	3.5	96.3	0.3	400
	Dutch	5.0	95.0	0.0	40
	Italian	2.7	97.3	0.0	73
	Brazilian	9.7	90.3	0.0	31
	Portuguese	6.3	87.5	6.3	16
	Turkish	4.8	95.2	0.0	21

Table 3: NMT quality experimental results.

The highest improvements were obtained for the PERSON category in Brazilian Portuguese with 98% of sentences showing better quality. In this particular case, the majority of these improvements are related to punctuation and register. For the other languages, the main difference was avoiding name omissions and hallucinations.

For the DATE category, the improvements were similar across all evaluated languages with gains up to 68.5% in the test cases. This shows that this category is prone to be distorted by the NMT. Looking at the sentences where it performed worse, a more in-depth analysis showed that the main issues were related to the translation of ordinal numbers, as well as the wrong preposition before the date, a consequence of using the generic entity token mask.

In what respects COUNTRY, it is possible to conclude that this is the category with the lowest percentage of improvements. The majority of sentences remained the same. This is because the entity fallback mechanism was not triggered often, which is in line with the fact that this is a NE with a limited number of realizations. This highlights the importance of entity fallback since otherwise, we could be introducing many agreement errors unnecessarily. In the few cases where the quality slightly decreased, the root cause was mainly the use of wrong prepositions before the NE when a valid translation did not match the dictionary.

5 Conclusions and Future Work

In this work, we presented a NE handling process, with the ultimate goal of bootstrapping an existing NLP pipeline to improve translation quality. This problem was tackled from a perspective of allowing such improvements to be delivered without having to change one of the main components of the pipeline, the NMT. By having this decoupling, the improvements can be delivered fast, enhancing the user experience in situations where NE translation errors can lead to catastrophic communication errors. Our process is based on dedicated recognition and NE translation steps. Integration into the existing NMT is done through semantic equivalent masking and an entity fallback mechanism. To evaluate NE recognition, we compared our domain custom model against two out-of-the-box models. The results show that the trade-off between recognition performance and data acquisition costs justifies a custom model for our use case. To evaluate our overall approach, we compared the translation quality of NE of the existing pipeline with the improved version. It was possible to observe that we achieved *translation quality* improvements while affording *translation guarantee* at the same time, validating our approach.

We also want to highlight that our approach allows us to easily anonymize Personally Identifiable Information (PII) data by exposing the NE mask rather than its original text. This is a concern for us since our NLP pipeline supports a feed-

back loop between NMT and human post-edition. The semantic equivalent mask is advantageous in this scenario since it allows editors to review more natural-looking sentences and without the cognitive overhead of processing a generic placeholder.

Regarding future work, one of the concerns is how to extend the generation of semantic equivalent NE to categories other than `PERSON`. The main obstacle is identifying the necessary linguistic properties for the generation in all necessary target languages. Another concern is the scalability of the NE recognition component. Thus far, our solution has been efficient since we have an overarching domain that ties in otherwise different topics. When moving to a completely different domain, we want to investigate how to keep this efficiency in collecting new data while leveraging the existing model.

6 Acknowledgements

This work was supported by national funds in Portugal through Fundação para Ciência e a Tecnologia (FCT), with reference UIDB/50021/2020 and through FCT and Agência Nacional de Inovação with the Project Multilingual AI Agents Assistants (MAIA), contracted number 045909.

References

Bordes, Antoine, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Proceedings of NIPS 2013, International Conference on Neural Information Processing Systems*, pages 2787–2795.

Chen, Yanqing and Steven Skiena. 2016. False-friend detection and entity matching via unsupervised transliteration. *arXiv preprint arXiv:1611.06722*.

Feng, Steven Y., Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Proceedings of ACL 2021, Annual Meeting of the Association for Computational Linguistics*, pages 968–988.

Hassan Awadalla, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, Will Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. arXiv:1803.05567.

Hochreiter, Sepp and Jürgen Schmidhuber. 1996. Lstm can solve hard long time lag problems. In *Proceedings of NIPS 1996, International Conference on Neural Information Processing Systems*, page 473–479.

Honnibal, Matthew, Ines Montani, Sofie Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python. To appear.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of EACL 2017, European Chapter of the Association for Computational Linguistics*, pages 427–431.

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

Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the ICML 2001, International Conference on Machine Learning*, pages 282–289.

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL 2016, North American Chapter of the Association for Computational Linguistics*, pages 260–270.

Li, Xiaoqing, Jiajun Zhang, and Chengqing Zong. 2016. Towards zero unknown word in neural machine translation. In *Proceedings of the IJCAI 2016, International Joint Conference on Artificial Intelligence*, page 2852–2858.

Li, Xiaoqing, Jinghui Yan, Jiajun Zhang, and Chengqing Zong. 2019. Neural name translation improves neural machine translation. In Chen, Jiajun and Jiajun Zhang, editors, *Proceedings of CWMT 2019, China Workshop on Machine Translation*, pages 93–100.

Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of ICCV 2017, International Conference on Computer Vision*, pages 2999–3007.

Merhav, Yuval and Stephen Ash. 2018. Design challenges in named entity transliteration. In *Proceedings of CLNLP 2018, International Conference on Computational Linguistics*, pages 630–640.

Modrzejewski, Maciej, Miriam Exel, Bianka Buschbeck, Thanh-Le Ha, and Alexander Waibel. 2020. Incorporating external annotation to improve named entity translation in NMT. In *Proceedings of EACL 2020, Annual Conference of the European Association for Machine Translation*.

Moussallem, Diego, Axel-Cyrille Ngonga Ngomo, Paul Buitelaar, and Mihael Arcan. 2019. Utilizing knowledge graphs for neural machine translation augmentation. In *Proceedings of K-CAP, International Conference on Knowledge Capture*, page 139–146.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014, Empirical Methods in Natural Language Processing*, pages 1532–1543.

Post, Matt, Shuoyang Ding, Marianna Martindale, and Winston Wu. 2019. An exploration of placeholding in neural machine translation. In *Proceedings of MT-Summit 2019, Machine Translation Summit*, pages 182–192.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the ACL 2020, Annual Meeting of the Association for Computational Linguistics*, pages 101–108.

Sennrich, Rico and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of WMT 2016, Conference on Machine Translation*, pages 83–91.

Shavarani, Hassan S. and Anoop Sarkar. 2021. Better neural machine translation by extracting linguistic information from BERT. In *Proceedings of EACL 2021, European Chapter of the Association for Computational Linguistics*, pages 2772–2783.

Talmor, Alon and Jonathan Berant. 2019. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of ACL 2019, Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of LREC 2012, International Conference on Language Resources and Evaluation*, pages 2214–2218.

Ugawa, Arata, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural machine translation incorporating named entity. In *Proceedings of CLNLP 2018, 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 *Proceedings of NIPS 2017, International Conference on Neural Information Processing Systems*, pages 5998–6008.

Vrandečić, Denny and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

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 WMT 2017, Conference on Machine Translation*, pages 410–415.

Wang, Xinyu, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of ACL 2021, Annual Meeting of the Association for Computational Linguistics*, pages 1800–1812.

Xie, Shufang, Yingce Xia, Lijun Wu, Yiqing Huang, Yang Fan, and Tao Qin. 2022. End-to-end entity-aware neural machine translation. *Machine Learning*, pages 1–23.

Zhao, Yang, Lu Xiang, Junnan Zhu, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2020a. Knowledge graph enhanced neural machine translation via multi-task learning on sub-entity granularity. In *Proceedings of CLNLP 2020, International Conference on Computational Linguistics*, pages 4495–4505.

Zhao, Yang, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2020b. Knowledge graphs enhanced neural machine translation. In *Proceedings of IJCAI 2020, International Joint Conference on Artificial Intelligence*, pages 4039–4045.

Zhou, Hao, Weidong Ren, Gongshen Liu, Bo Su, and Wei Lu. 2021. Entity-aware abstractive multi-document summarization. In *Proceedings of ACL 2021, Annual Meeting of the Association for Computational Linguistics*, pages 351–362.