Look It Up: Bilingual and Monolingual Dictionaries Improve Neural Machine Translation

Xing Jie Zhong
University of Notre Dame
xzhong3@nd.edu

David Chiang
University of Notre Dame
dchiang@nd.edu

Abstract

Despite advances in neural machine translation (NMT) quality, rare words continue to be problematic. For humans, the solution to the rare-word problem has long been dictionaries, but dictionaries cannot be straightforwardly incorporated into NMT. In this paper, we describe a new method for "attaching" dictionary definitions to rare words so that the network can learn the best way to use them. We demonstrate improvements of up to 3.1 BLEU using bilingual dictionaries and up to 0.7 BLEU using monolingual source-language dictionaries.

1 Introduction

Despite its successes, neural machine translation (NMT) still has unresolved problems. Among them is the problem of rare words, which are paradoxically very common because of Zipf's Law. In part, this is a problem intrinsic to data-driven machine translation because the system will inevitably encounter words not seen in the training data. In part, however, NMT systems seem particularly challenged by rare words, compared with older statistical models.

One reason is that NMT systems have a fixed-size vocabulary, typically 10k–100k words; words outside this vocabulary are represented using a special symbol like UNK. Byte pair encoding (BPE) breaks rare words into smaller, more frequent subwords, at least allowing NMT to see them instead of UNK (Sennrich et al., 2016). But this by no means solves the problem; even with subwords, NMT seems to have difficulty learning translations of very rare words, possibly an instance of catastrophic forgetting (McCloskey and Cohen, 1989).

Humans deal with rare words by looking them up in a dictionary, and the idea of using dictionaries to assist machine translation is extremely old. From a statistical perspective, dictionaries are a useful complement to running text because the uniform distribution of dictionary headwords can smooth out the long-tailed distribution of running text. In pre-neural statistical machine translation systems, the typical way to incorporate bilingual dictionaries is simply to include them as parallel sentences in the training data. But (as we show), this does not work well for NMT systems.

We are aware of only a few previous attempts to find better ways to incorporate bilingual dictionaries in NMT. Some methods use dictionaries to synthesize new training examples (Zhang and Zong, 2016; Qi et al., 2018; Hämäläinen and Alnajjar, 2019). Arthur et al. (2016) extend the model to encourage it to generate translations from the (automatically extracted) dictionary. Post and Vilar (2018) constrain the decoder to generate translations from the dictionary. What these approaches have in common is that they all treat dictionary definitions as target-language text, when, in fact, they often have properties very different from ordinary text. For example, CEDICT defines 此致 (*cǐzhì*) as "(used at the end of a letter to introduce a polite salutation)" which cannot be used as a translation. In the case of a monolingual source-language dictionary, the definitions are, of course, not written in the target language at all.

In this paper, we present an extension of the Transformer (Vaswani et al., 2017) that "attaches" the dictionary definitions of rare words to their occurrences in source sentences. We introduce new position encodings to represent the nonlinear structure of a source sentence with its attachments. Then the unmodified translation model can learn how to make use of this attached information. We show that this additional information yields improvements in translation accuracy of up to 3.1 BLEU. Because our method does not force dictionary definitions to be treated as target-language text, it is generalizable to other kinds of information, such as monolingual source-language dictionaries, which yield smaller improvements, but still as much as 0.7 BLEU.

Proceedings of the 5th Conference on Machine Translation (WMT), pages 538–549
Online, November 19–20, 2020.

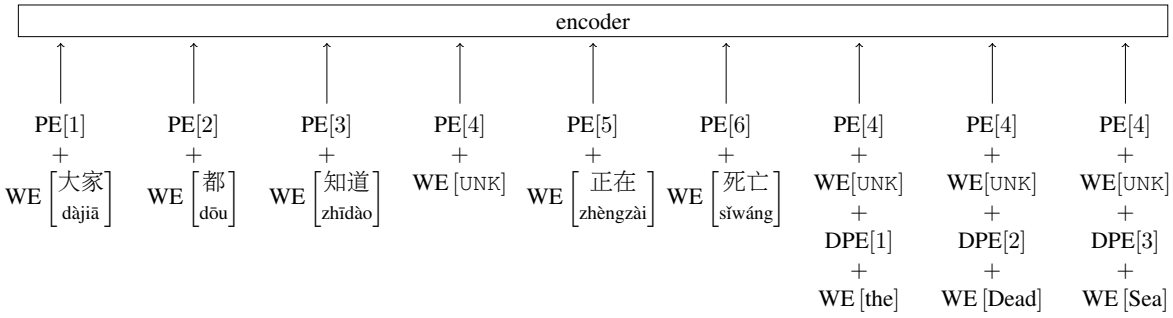

Figure 1: Our method attaches dictionary definitions to rare words. Here, the source sentence is 大家 都 知道 死海 正在 死亡 (*dàjiā dōu zhīdào Sǐhǎi zhèngzài sǐwáng*, *Everyone knows that the Dead Sea is dying*). WE[f] is the embedding of word f, PE[p] is the encoding of position p, and DPE[q] is the encoding of position q within a dictionary definition. The rare word 死海 (*Sǐhǎi*) is replaced with UNK and defined as *the Dead Sea*. The words of the definition are encoded with both the position of the defined word (4) and their positions within the definition.

2 Methods

Our method is built on top of the Transformer (Vaswani et al., 2017). For each unknown source word with an entry in the dictionary, we attach the first 50 tokens of the definition (discarding the rest of the definition) to the source sentence. As described below, we encode the definition so as to differentiate it from the source sentence proper and to record which source word the definition is attached to. We leave the task of deciding whether and how to use the definition up to the translation model, which we use without any modifications.

2.1 Position encodings

To differentiate the attached definitions from the source sentence itself, we use special position encodings.

An ordinary word f at position p is encoded, as usual, as $\mathrm{E}[f] = \mathrm{WE}[f] + \mathrm{PE}[p]$, where WE is the word embedding and PE is the usual sinusoidal position encoding (Vaswani et al., 2017).

Suppose that word f at position p has an attached definition. Then word d at position q of the definition is encoded as

$$\mathrm{E}[d] = \mathrm{WE}[f] + \mathrm{PE}[p] + \mathrm{WE}[d] + \mathrm{DPE}[q],$$

where DPE is a position encoding scheme different from PE. We experimented with several schemes for DPE; in the experiments below, we learned a different encoding for each position (Gehring et al., 2017).

See Figure 1 for an illustration of the encoding of an example source sentence. Note that once all words have received their position encodings, their order does not matter, as the Transformer encoder is order-independent.

2.2 Subword segmentation

To apply our method on data that has been segmented using BPE, we face two new problems. First, since very few words are replaced with UNK, it is not sufficient only to attach definitions to UNK. How do we decide which words to attach definitions to? Second, if a word has been split into multiple subwords, the definition does not have a single attachment position. How do we represent the attachment position when encoding the definition?

To choose which words to define, we use a simple frequency threshold. We scan the data (after tokenization/segmentation but before BPE) for matches with the dictionary, including multi-word matches. If any substring of the source sentence matches a headword in the dictionary and occurs in the training data k or fewer times, we attach its definition. The threshold k can be tuned on the development data.

To attach a definition to a substring with more than one token, we simply fuse all the tokens in the substring into a single token, which often (but not always) then falls out of the vocabulary and is therefore changed to UNK. We attach the dictionary definition to this single token, which represents the whole word or expression.

For example, in the sentence in Figure 1, BPE splits 死海 (*sǐhǎi*) into 死@@ 海 (*sǐ@@ hǎi*) (where @@ is the marker that typical implementations of BPE use to indicate subword splits). Assuming that 死海 occurs k or fewer times, we fuse it back into a single token, which gets changed into UNK. Then the dictionary definition is attached as described above.

Language	Task	lines				words		
		train	dev	test	total	tokens	types	vocab
Chi-Eng	Spoken	176,000	22,000	22,000	220k	5.9M	179k	25k
	Science	216,000	27,000	27,000	270k	10.1M	383k	27k
	Laws	176,000	22,000	22,000	220k	17.4M	98k	22k
	News	360,000	45,000	45,000	450k	25.3M	477k	24k
	Education	360,000	45,000	45,000	450k	18.6M	461k	28k
	Subtitles	240,000	30,000	30,000	300k	6.6M	147k	27k
	Thesis	240,000	30,000	30,000	300k	17.2M	613k	27k
	UM-all	1,993,500	221,500	5,000	2.2M	101.3M	1.3M	33k
Deu-Eng	Europarl-small	160,000	20,000	20,000	200k	10.9M	151k	16k
	Europarl-all	1,440,000	180,000	197,758	1.8M	98.6M	475k	16k

Table 1: Statistics of the various tasks we experimented on. Train/dev/test: number of lines selected for use as training, development, and test data (respectively). Toks: number of word tokens (source+target). Types: number of word types (source+target). Vocab: joint vocabulary size used in word-based experiments.

3 Experiments

In this section, we describe our experiments on Chinese-English and German-English translation, comparing our method (which we call *Attach*) against two baselines. One baseline is the standard Transformer without any dictionary information (which we call *Baseline*). The other baseline is the standard Transformer with the bilingual dictionaries included as parallel sentences in the training data (which we call *Append*).

3.1 Data: Chinese-English

For Chinese-English, we used the UM-Corpus[1] (Tian et al., 2014), which has about 2M sentence pairs in eight different domains. Since rare words may be more frequent in certain domains, testing our model on different types of data may highlight the conditions where dictionaries can be helpful. We excluded the Microblog domain because of its length (only 5000 lines). For each of the other domains, we split the data into three parts: the first roughly 80% for training (*train*), the next 10% for development (*dev*), and the last 10% for testing (*test*). The task *UM-all* combines all eight domains. The UM-Corpus provides a test set, which we used (*test*), and we split the provided training data into two parts, the first 90% for training (*train*) and last 10% for development (*dev*). The exact line counts and other statistics are shown in Table 1.

We used the Stanford segmenter[2] (Chang et al., 2008) for the Chinese data and the Moses tokenizer[3] for the English data.

As a dictionary, we used CC-CEDICT,[4] which has 116,493 entries. Each entry has a traditional Chinese headword (which we delete), a simplified Chinese headword, a pronunciation (which we delete), and one or more definitions. We process the definitions as follows:

- Remove substrings of the form *abbr. for* c, where c is a Chinese word.
- If a definition contains *see* c or *see also* c, where c is a Chinese word, replace it with the definition of c.
- Remove everything in parentheses.
- Remove duplicate definitions.
- If the entry has no definitions left, delete the whole entry.
- Concatenate all the definitions into a single string.

The resulting dictionary has 102,567 entries, each consisting of a Chinese headword and a single English definition. We segmented/tokenized these in the same way as the parallel data. The average definition length is five, and the maximum definition length is 107.

[1] http://nlp2ct.cis.umac.mo/um-corpus/

[2] https://nlp.stanford.edu/software/segmenter.shtml

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

[4] https://www.mdbg.net/chinese/dictionary?page=cedict, downloaded 10/2018.

For example, consider the following CEDICT entries, where we have already removed traditional Chinese characters and pronunciations for clarity.

三自 /abbr. for 三自爱国教会, Three-Self Patriotic Movement/
U盘 /USB flash drive/see also 闪存盘
闪存盘 /USB flash drive/jump drive/thumb drive/memory stick/

After cleaning, these would become

三自 Three-Self Patriotic Movement
U盘 USB flash drive jump drive thumb drive memory stick
闪存盘 USB flash drive jump drive thumb drive memory stick

3.2 Data: German-English

For German-English, we used the Europarl V7 dataset.[5] We tokenized both sides of the data with the Moses tokenizer. Due to the size of the original Europarl dataset and the increased runtime from our method, we ran some experiments on only the first 200k lines of the dataset, denoted in result tables as *Europarl-small*, while the full Europarl data is called *Europarl-all*. We split both into three parts: the first roughly 80% for training, the next 10% for development, and the last 10% for testing. Some statistics of the data are shown in Table 1.

We used the German-English dictionary from Stardict,[6] which is derived from Freedict[7] and has 81,628 entries. In this dictionary, the headwords have notes in parentheses indicating things like selectional restrictions; we deleted all of these. Unlike with CEDICT, we did not delete any material in definitions, nor did we resolve cross-references, which were very rare. As before, we removed blank entries and merged multiple definitions into a single line. We tokenized both headwords and definitions with the Moses tokenizer. The final dictionary size is 80,737 entries, with an average definition length of 2.9 and a maximum definition length of 88.

For example, the entry:

(Aktien) zusammenlegen to merge (with)

would become

zusammenlegen to merge (with)

[5] http://statmt.org/europarl/
[6] http://download.huzheng.org/freedict.de/
[7] https://freedict.org/

Task	Baseline	Append	Attach
Spoken	13.6	12.4	15.4
Science	8.0	6.6	9.2
Laws	29.0	27.4	30.2
News	9.9	10.2	11.2
Education	9.1	8.7	9.9
Subtitles	18.3	16.4	20.2
Thesis	9.5	9.5	10.6
UM-all	16.8	16.7	17.7
Europarl-small	29.2	28.4	29.6
Europarl-all	30.0	29.8	30.1

Table 2: Results on word-based translation. Our method (Attach) significantly improves over the baseline in all tasks. Appending the dictionary to the parallel data (Append) performs worse in all tasks except in News; differences are significant for all tasks except UM-all and Thesis.

3.3 Implementation and details

We used Witwicky,[8] an open-source implementation of the Transformer, with all of its default hyperparameters. We use the same random seed in each experiment. We modified it to attach dictionary definitions as described above. The code and our cleaned dictionaries are available under an open-source license.[9]

For BPE-based translation, we used joint BPE with 16k operations. For word-based translation, we set each system's vocabulary size close to the vocabulary size of the corresponding BPE-based system. For example, the Spoken dataset with 16k BPE applied to the training data has 25,168 word types, so we limited the word-based model to 25,000 word types. The vocabulary size we chose for each data set is shown in Table 1.

For all tasks except UM-all and Europarl-all, we trained for 20 epochs, and used the model with the highest dev BLEU to translate the test set. Due to the massive increase in training data on the UM-all and Europarl-all datasets, we only trained for 10 epochs. Otherwise, the settings are the same across all experiments.

We report case-insensitive BLEU scores of detokenized outputs against raw references. We perform significance testing with bootstrap resampling using 1000 samples, with a significance level of 0.05.

[8] https://github.com/tnq177/witwicky
[9] https://github.com/xjz92/Attach-Dictionary

Method	UM-Spoken Dev BLEU
Baseline	13.6
Attach to unknown words	13.9
+ fused multi-word expressions	13.8
+ all words	13.8

Table 3: Comparison of variations of our method on word-based translation.

Method	UM-Spoken Dev BLEU
Baseline	14.2
Attach to fused unknown words	14.8
+ fused multi-word expressions	14.8

Table 4: Comparison of variations of our method on BPE-based translation.

3.4 Results: Word-Based

Table 2 shows results on word-based translation. The *Append* column shows that simply appending the bilingual dictionary to the parallel training data is not helpful, for all tasks except News; these differences are significant for all tasks except UM-all and Thesis. By contrast, our method improves accuracy significantly over the baseline across all tasks.

We also compared against some variations of our method. First, CEDICT has definitions for single words as well as multi-word expressions. In our original setup, we only look up unknown single words, so the definitions for multi-word expressions are never used. To fully utilize the dictionary, we tried changing the source data by taking every substring that matches a dictionary entry and fusing it into a single token, which would often, but not always, fall out of the vocabulary and be changed to `UNK`. When more than one match was possible, we chose the longest possible match, breaking ties arbitrarily. However, we found that fusing phrases did not perform as well as just fusing words (Table 3). We also tried attaching dictionary definitions to all tokens, not just `UNK` tokens. Unfortunately, this also did not perform as well (Table 3).

3.5 Results: BPE-Based

As described in Section 2.2, we fuse subwords in order to attach definitions. Again we must first decide whether we wanted to fuse multi-word expressions.

Task	Baseline	Append	Fuse	Attach
Spoken	16.6	14.7	16.3	17.0
Science	11.6	9.6	13.8	14.7
Laws	29.0	26.8	29.0	30.0
News	11.8	10.9	11.3	13.3
Education	12.9	12.3	12.2	14.2
Subtitles	20.0	17.3	19.7	21.3
Thesis	15.3	14.2	14.9	15.5
UM-all	19.8	19.7	19.3	21.4
Europarl-small	32.6	30.8	33.4	33.5
Europarl-all	35.3	36.0	36.1	36.5

Table 5: Results on BPE-based translation. Our method (Attach) improves significantly over the baseline in Europarl-small and all Chinese-English tasks, whereas appending the dictionary to the parallel data (Append) performs worse, significantly so for Europarl-small and all Chinese-English tasks except UM-all. For Europarl-all, Append is significantly better. The Fuse column shows the effect of fusing words that would receive definitions, without actually attaching the definitions.

On the dev set, both methods have comparable performance (Table 4). Since we were interested in using as much of the dictionary as possible, we chose the model that fuses phrases.

As described in Section 2.2, we fuse subwords and attach definitions only for words whose frequency falls below a threshold. To tune this threshold, we trained models using thresholds of $k = 5$, 10, 15, 20, 50, 100, and ∞, and measured BLEU on the development set (Figure 2). We found that for Chinese-English, $k = \infty$ was best, but for German-English, $k = 5$ was best.

The results are shown in Table 5. As before, we compared against the two baselines (*Baseline* and *Append*). To tease apart the effect of fusing words and adding dictionary definitions, we also tested a model where all words that would receive definitions are fused, but the definitions are not actually attached (*Fuse*). Finally, we tested our model (*Attach*). On Chinese-English, our model improved significantly over the baselines across all tasks, whereas appending the dictionary to the parallel data did worse, significantly so on all tasks except UM-all. On German-English, the results on Europarl-small were similar, with Append doing significantly worse and our model doing significantly better. Interestingly, on Europarl-all, Append does significantly better than the baseline.

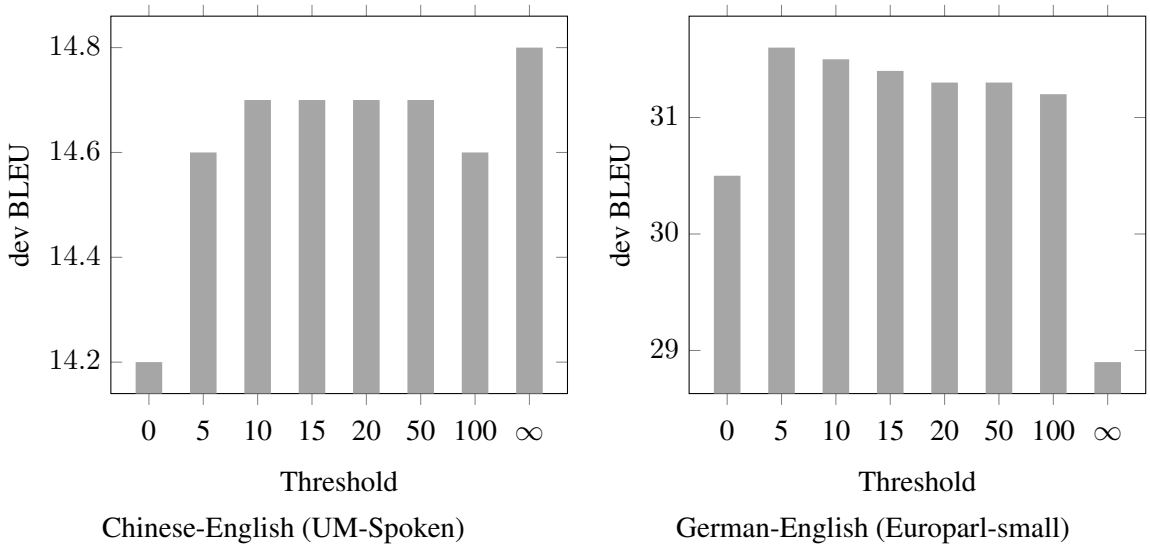

Figure 2: Effect on dev BLEU scores of the frequency threshold below which we fuse a word and attach its definition. These scores are used to choose the threshold that is used in Table 5.

3.6 Monolingual dictionaries

Because our dictionary-attachment method does not make any assumptions about the form of the definitions, we can apply it to monolingual source-language dictionaries as well. Monolingual source-language dictionaries are a natural resource for human translators, but we're not aware of previous research that uses them in data-driven machine translation. For many languages and language pairs, we expect them to be much more comprehensive than bilingual dictionaries. Our monolingual dictionary is the 汉语辞海 (*Hànyǔ Cíhǎi*),[10] which has a total of 380,579 entries. We removed pronunciations and concatenated multiple definitions into a single line. We did not resolve any cross-references in this dictionary, and we removed all entries with empty definitions. This gives us a final dictionary size of 358,234 entries.

We experimented with using this dictionary on the Spoken and Science UM datasets. The results are shown in Table 6. Although, as expected, it does not help as much as a bilingual dictionary, it does help on three out of four tasks we tried. All differences in this table are statistically significant.

		Test BLEU	
Segmentation	Dictionary	Spoken	Science
word	none	13.6	8.0
	zh–zh	14.3	8.4
	zh–en	15.4	9.2
BPE	none	16.6	11.2
	zh–zh	15.2	11.6
	zh–en	17.0	14.7

Table 6: Attaching a monolingual Chinese-Chinese dictionary improves over the baseline in three out of four tasks, although not as much as a bilingual Chinese-English dictionary does. All differences are statistically significant.

4 Analysis

To further examine how our methods improve translation, we looked at some examples in our UM-Spoken dev set, shown in Table 7 (word-based) and Table 8 (BPE). The (UNK) tag next to dictionary definitions indicates that the word is outside of the system's vocabulary.

In the first example, 对称性 (*duìchènxìng*, symmetry) is unknown to the word-based systems. Adding the definition to the parallel training data (*Append*) does not help word-based translation because the word remains unknown, whereas our model correctly generates the translation *symmetry*. With BPE, the word is broken into three pieces, so that the Append system can correctly generate the word *symmetry*. But the third character (性, *xìng*) can also mean "sex," and together with the following character (性感, *xìnggǎn*) can mean "sexy." This explains why the Append system incorrectly adds the words *of sex*.

In the second example, 火药 (*huǒyào*, gunpow-

[10] http://download.huzheng.org/zh_CN/

Source	1. 不 只 是 科学家们 对 对称性(UNK) 感 兴趣 。 2. 我 哥哥 听说 我们 做 了 火药(UNK) 。 3. 有些 登山者 经过 他 身旁 ， 打量(UNK) 了 他 一 番
Definitions	1. 对称性: symmetry 2. 火药: gunpowder(UNK) 3. 打量: to size sb(UNK) up to look sb(UNK) up and down to take the measure of to suppose to reckon
Reference	1. But it's not just scientists who are interested in symmetry. 2. Well, my brother heard that we had made gunpowder. 3. Some climbers had come by and looked at him,
Baseline	1. not only scientists are interested in the UNK of UNK. 2. My brother has heard that we've done a lot of work. 3. And some of the climber went to him, and he said,
Append	1. It's not just about scientists who are interested in UNK. 2. My brother has heard that we've done a lot of work. 3. And some of the UNK came over and over and over again,
Attach	1. not just scientists are interested in symmetry. 2. My brother heard that we had done UNK. 3. Some of the climber passed him, looked at him,

Table 7: Examples from word-based systems on the UM-Spoken data. In the first and second examples, the unknown words 对称性 (*duìchènxìng*) and 火药 (*huǒyào*) cannot be translated by the baseline, even with the dictionary in the parallel data (Append). Our model successfully incorporates the dictionary definition *symmetry*, but not *gunpowder*, because it is unknown. In the third example, the definition is not suitable as a direct translation of the unknown word 打量 (*dǎliàng*), but our model generates the word *looked*, apparently by picking out the word *look* from the definition and inflecting it correctly for the context.

BPE Source	1. 不 只 是 科学家们 对 对@@ 称@@ 性 感 兴趣。 2. 我 哥哥 听说 我们 做 了 火@@ 药 。 3. 有些 登@@ 山@@ 者 经过 他 身@@ 旁 ， 打@@ 量 了 他 一 番
Fused source	1. 不 只 是 科学家们 对 对称性(UNK) 感 兴趣 。 2. 我 哥哥 听说 我们 做 了 火药(UNK) 。 3. 有些 登山@@ 者 经过 他 身@@ 旁 ， 打量(UNK) 了 他 一 番 ,
Definitions	1. 对称性: sym@@ metry 2. 火药: gun@@ powder 3. 打量: to size s@@ b up to look s@@ b up and down to take the measure of to suppose to reck@@ on
Reference	1. But it's not just scientists who are interested in symmetry. 2. Well, my brother heard that we had made gunpowder. 3. Some climbers had come by and looked at him,
Baseline	1. not just scientists are interested in the sense of sympathy. 2. My brother had heard that we did a fire pills. 3. Some of the climbers passed him on the side, and he had a lot of money,
Append	1. Not only scientists are interested in the symmetry of sex. 2. My brother told us that we had done a fire. 3. Some of the climber passed his feet, and he took a second,
Fuse	1. not only scientists are interested in their interests in the world. 2. My brother has heard that we've done a good job. 3. Some of the climbers passed by him, and he gave him a sense,
Attach	1. not only scientists are interested in symmetry. 2. My brother heard that we did the gunpowder. 3. Some climbers passed by his side and looked at him,

Table 8: Examples from BPE-based systems on the UM-Spoken data. In the first two examples, the baseline system, even with the dictionary in the parallel data (Append), tries to translate the pieces of unknown words separately and incorrectly (e.g., *fire*, *pills*, *sex*). Our model is able to translate the first and third examples correctly as in Table 7, as well as the second example.

der) is unknown, and the definition word *gunpowder* is also unknown. So none of the systems are able to translate this word correctly (though arguably our system's generation of UNK is preferable). When we switch to BPE, our model generates the correct translation. The other systems fail because this word splits into two very common words, 火 (*hǔo*, fire), and 药 (*yào*, medicine), which the system tries to translate separately.

The third example shows what happens when we have a long definition that contains useful information, but is not suitable as a direct translation of the unknown word 打量 (*dǎliàng*). Here we see that our attachment model generates the word *looked*, apparently by picking out the word *look* from the definition and inflecting it correctly for the context. No other models were able to generate a word with a similar meaning.

Please see Appendix A for visualizations of the encoder-decoder attention for these three examples.

We also looked at a few examples from the Europarl-small dev set, shown in Table 9 and 10. In the first example, the definition *omission* was out of vocabulary, so our model was not able to perform any better than the baselines. However, in the BPE systems, our model was able to properly translate *Auslassung* to *omission* while none of the other baseline systems was able to. In the second example, we see something similar in the word-based system. The Baseline and Append models were unable to generate the correct translation of *Alternativlösung*, but our method was. With BPE, all systems (even Baseline) were able to translate the word correctly.

5 Discussion

In Section 1, we mentioned several other methods for using dictionaries in NMT, all of which treat dictionary definitions as target-language text. An alternative approach to handling rare words, which avoids dictionaries altogether, is to use word embeddings trained on large amounts of monolingual data, like fastText embeddings (Bojanowski et al., 2017). Qi et al. (2018) find that fastText embeddings can improve NMT, but there is a sweet spot (likely between 5k and 200k lines) where they have the most impact. They also find that pre-trained embeddings are more effective when the source and target languages are similar.

We, too, experimented with using fastText word embeddings in our NMT system, but have not seen any improvements over the baseline – perhaps because our datasets are somewhat larger than those used by Qi et al. (2018). We also experimented with using dictionaries to improve word embeddings and found that the present approach, which gives the model direct access to dictionary definitions, is far more effective.

The most significant limitation of our method is runtime: because it increases the length of the source sentences, training and decoding take 2–3 times longer. Another limitation is that the effectiveness of this method depends on the quality and coverage of the dictionaries.

In the future, we plan to experiment with additional resources, like thesauruses, gazetteers, or bilingual dictionaries with a different target language. Second, from our examples, we see that our model is able to select a snippet of the definition and adapt it to the target context (for example, by inflecting words), but further analysis is required to understand how much the model is able to do this. Finally, our method currently requires an exact match between a dictionary headword and a source word; we plan to extend the model to enable matching of headwords with inflected forms.

6 Conclusion

In this paper, we presented a simple yet effective way to incorporate dictionaries into a Transformer NMT system, by attaching definitions to source sentences to form a nonlinear structure that the Transformer can learn how to use. We showed that our method can beat baselines significantly, by up to 3.1 BLEU. We also analyzed our system's outputs and found that our model is learning to select and adapt parts of the definition, which it does not learn to do when the dictionary is simply appended to the training data. We also found that our method has some potential to work with monolingual dictionaries.

Acknowledgements

This paper is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract #FA8650-17-C-9116. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Gov-

Source	1. Ich hoffe , dass diese Auslassung(UNK) korrigiert werden kann . 2. Wäre das nicht eine Alternativlösung(UNK) ?
Definitions	1. Auslassung: omission(UNK) 2. Alternativlösung: alternative solution
Reference	1. I hope that this omission can be corrected. 2. Would this not be an alternative solution?
Baseline	1. I hope that this UNK can be corrected. 2. Would this not be a UNK?
Append	1. I hope that this UNK can be corrected. 2. Would this not be a UNK?
Attach	1. I hope that this UNK can be corrected. 2. Would this not be an alternative solution?

Table 9: Examples from word-based systems run on the Europarl-small data. In the first example, the dictionary defines unknown word *Auslassung* with another unknown word, *omission*, so neither adding the dictionary to the parallel data (Append) nor our model (Attach) benefits. In the second example, adding the dictionary definition of *Alternativlösung* to the parallel data does not help, but our model is able to incorporate it.

BPE source	1. Ich hoffe , dass diese Aus@@ l@@ assung korrigi@@ ert werden kann . 2. W@@ äre das nicht eine Altern@@ ativ@@ lösung ?
Fused source	1. Ich hoffe , dass diese Auslassung(UNK) korrigi@@ ert werden kann . 2. W@@ äre das nicht eine Alternativlösung(UNK) ?
Definitions	1. Auslassung: om@@ is@@ sion 2. Alternativlösung: alternative solution
Reference	1. I hope that this omission can be corrected. 2. Would this not be an alternative solution?
Baseline	1. I hope that this approval can be corrected. 2. Would this not be a alternative solution?
Append	1. I hope that this interpretation can be corrected. 2. Would this not be a alternative solution?
Fuse	1. I hope that these pieces can be corrected. 2. Would this not be a pronounce?
Attach	1. I hope that this omission can be corrected. 2. Would this not be an alternative solution?

Table 10: Examples from BPE-based systems run on the Europarl-small data. In the first exapmle, unlike in Table 9, the unknown word *Auslassung* is not replaced with UNK but is split into subwords, which the baseline system as well as the system with the dictionary in its parallel data (Append) translate incorrectly. Our model successfully uses the dictionary definition, *omission*. In the second example, BPE enables all models to translate the compound *Alternatvlösung* correctly.

ernment is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proc. EMNLP*, pages 1557–1567.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Trans. ACL*, 5:135–146.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proc. WMT*, pages 224–232.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proc. ICML*, pages 1243–1252.

Mika Hämäläinen and Khalid Alnajjar. 2019. A template based approach for training NMT for low-resource Uralic languages - a pilot with Finnish. In *Proc. 2nd International Conference on Algorithms, Computing and Artificial Intelligence (ACAI)*, pages 520–525.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proc. NAACL HLT*, pages 1314–1324.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proc. NAACL HLT*, pages 529–535.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. ACL*, pages 1715–1725.

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, Yi Lu, Shuo Li, Yiming Wang, and Longyue Wang. 2014. UM-corpus: A large English-Chinese parallel corpus for statistical machine translation. In *Proc. LREC*, pages 1837–1842.

Ashish Vaswani, 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 30*, pages 5998–6008.

Jiajun Zhang and Chengqing Zong. 2016. Bridging neural machine translation and bilingual dictionaries. arXiv:1610.07272.

A Attention Visualizations

Figures 3 and 4 show visualizations of the attention of our Attach model. They show the first layer of encoder-decoder attention when translating the three Chinese sentences of Tables 7 and 8. Note the translations are not exactly the same as shown above, because we used a beam size of one instead of the default of four.

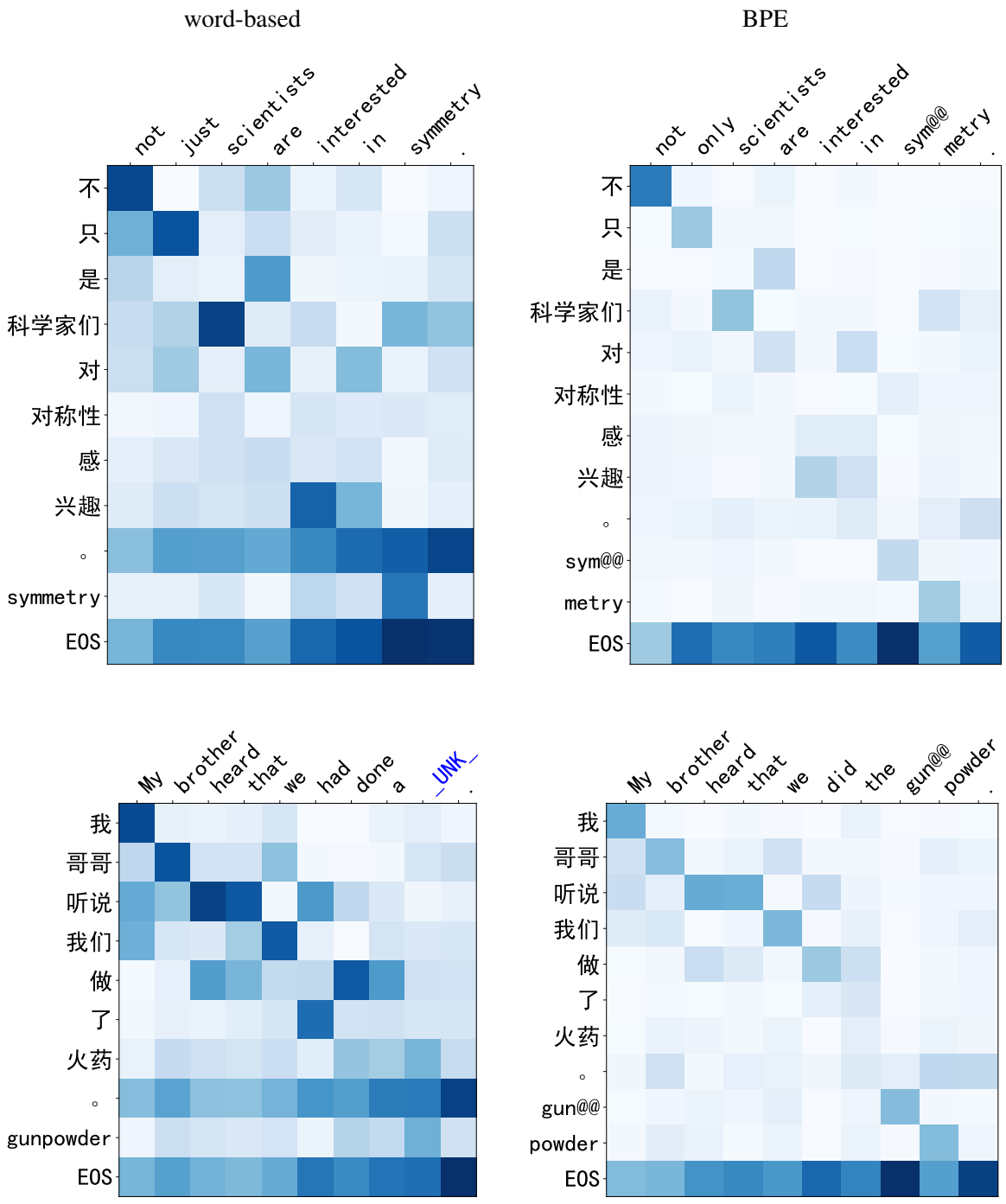

Figure 3: Attention visualizations for the first two Chinese-English examples of Tables 7 and 8.

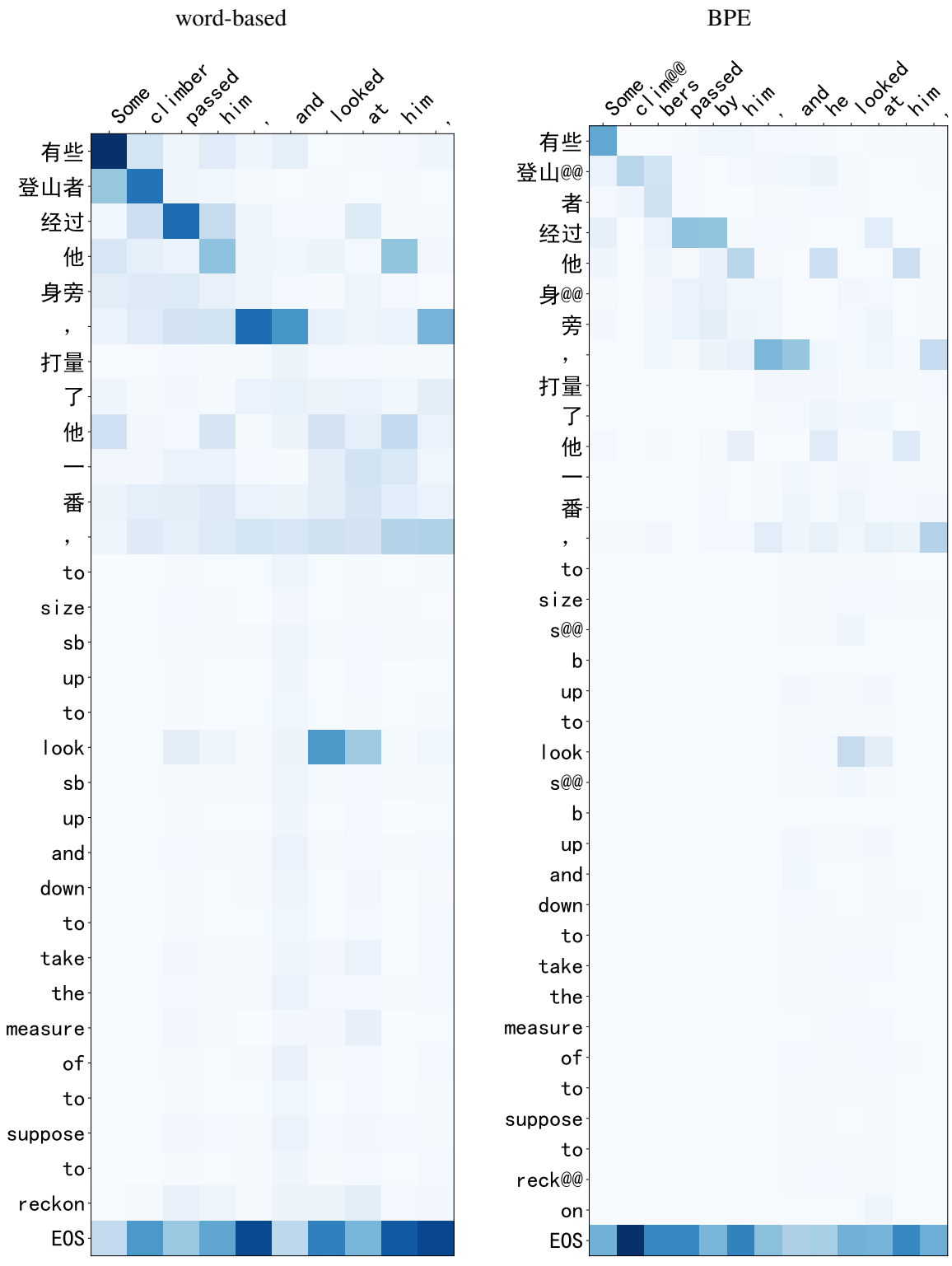

Figure 4: Attention visualizations for the third Chinese-English example of Tables 7 and 8.