Improving Rare Word Translation With Dictionaries and Attention Masking

Kenneth J. Sible ksible@nd.edu
David Chiang dchiang@nd.edu
Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, 46556, United States

Abstract

In machine translation, rare words continue to be a problem for the dominant encoder-decoder architecture, especially in low-resource and out-of-domain translation settings. Human translators solve this problem with monolingual or bilingual dictionaries. In this paper, we propose appending definitions from a bilingual dictionary to source sentences and using *attention masking* to link together rare words with their definitions. We find that including definitions for rare words improves performance by up to 1.0 BLEU and 1.6 MacroF1.

1 Introduction

The current state-of-the-art for machine translation (MT) is still the transformer encoder-decoder architecture (Kocmi et al., 2023). While large language models such as LLaMA and GPT-4 have achieved great success on various NLP tasks, they still fall behind dedicated encoder-decoders for MT (Xu et al., 2024). A major drawback of encoder-decoder models, however, is that they continue to struggle with rare word translation (Minh-Cong et al., 2022).

Dictionaries, both monolingual and bilingual, are an indispensable resource for human translators, and in pre-neural statistical MT systems, it was common to use bilingual dictionaries to improve translation of rare words (Tan et al., 2015). However, the use of dictionaries in neural MT is not straightforward, as there is a strong dependence on the surrounding context and word frequency in the training data (Wu et al., 2021). In this paper, we explore a new approach for incorporating dictionaries into neural MT systems. We hypothesize that dictionaries could be useful both for low-resource translation, where the target language has limited training data, and out-of-domain translation, where the testing domain differs significantly from the training domain(s). In addition, dictionaries could facilitate continual learning by enabling zero-shot adaptation of MT systems.

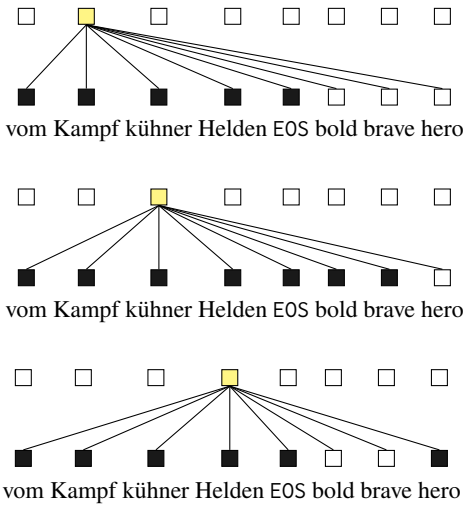

Figure 1: We append definitions of *kühn* 'bold, brave' and *Held* 'hero' to a sentence, and use an attention mask (with learnable strength) to inform the model which definitions correspond to which words. In each picture, the query vectors are above, with one query vector shaded yellow, and the key/value vectors are below, shaded to indicate the strength of the attention mask (black = not masked, white = masked).

Proceedings of the 16th Conference of the Association for Machine Translation in the Americas,
Chicago, USA, September 30 - October 2, 2024. Volume 1: Research Papers

However, the morphology of both the source and target language poses a major challenge for the use of dictionaries with MT compared to other NLP tasks that incorporate methods of retrieval-augmented generation (Niehues, 2021). MT systems that incorporate dictionaries must be capable of inflecting definitions for the target language and the context in which those definitions must appear, as dictionary entries and their definitions are often base forms. In Figure 1, we see the adjective *kühner* is declined for the genitive case, but only the lemma *kühn* would be in a German dictionary. Moreover, if the target language has adjective declension, then the MT system must also decline the dictionary form of the definition.

Our approach is to retrieve dictionary definitions for low-frequency words, append the definitions to source sentences containing rare words, and use attention masking to link together rare words with their definitions. We find that appending definitions for rare words improves MT performance by up to 1.0 BLEU and 1.6 MacroF1.

2 Related Work

Previous work on dictionaries for neural MT can be divided into two broad categories, which we call *dictionaries-as-translators* and *dictionaries-as-text*. In the dictionaries-as-translators approach, the dictionary is assumed to contain high-quality translations of words, and the technical challenge is to get the MT system to use the dictionary's translations when appropriate. In the dictionaries-as-text approach, dictionary entries are added somehow to the source sentence, and it is up to the MT system to learn how to use them. In this approach, the dictionary can contain definitions that are not necessarily translations (e.g., one definition for German *halt* is: "Indicating that something is generally known, or cannot be changed, or the like; often untranslatable"). This approach could, in principle, use other resources like monolingual dictionaries, grammars, and so on.

2.1 Dictionaries as translators

In the dictionaries-as-translators category, Zhang et al. (2021) propose a model with three steps: (1) identify source words that can be translated using a dictionary, (2) select one of several translation candidates (*i.e.*, definitions), and (3) copy the selected translation into the output sequence. Similarly, other previous work in this category uses constrained decoding with a translation lexicon: Zhang and Zong (2016), Arthur et al. (2016), Fadaee et al. (2017), Chatterjee et al. (2017), Hasler et al. (2018), Post and Vilar (2018), Thompson et al. (2019), Dinu et al. (2019).

A translation lexicon is a mapping of words from the source language to the target language, whereas a bilingual dictionary provides several possible translations for a given source word in addition to including definitions for untranslatable words such as particles. To incorporate a translation lexicon, we must constrain the output of the MT system, but that approach assumes the correct translation given the source context is contained within the lexicon. However, it quite often is the case that there are several valid translations with some being more appropriate than others for the given context.

2.2 Dictionaries as text

In the dictionaries-as-text category are approaches in which dictionary definitions are added to source sentences so that the model can learn how to use them. Two further questions arise: (1) How do we decide which definitions to include (especially in morphologically-rich languages, where a word in context does not in general match a dictionary headword)? (2) How do we represent the nonlinear structure of the input, which includes both a source sentence and associated definitions?

Niehues (2021) lemmatizes each rare word and retrieves the matching bilingual definition, if any. The definition is inserted into the sentence immediately after the rare word, delimited by #. He uses a combination of subword and character tokenization to improve handling of rare inflected forms.

Zhong and Chiang (2022) use a combination of Levenshtein distance and locality-sensitive hashing to find the closest dictionary headword for each, potentially inflected, rare word. They append the definitions to the end of the source sentence, and they inform the model about the structure of the input using position encodings (PEs). Each definition word's vector has contributions from both its own (sinusoidal) PE as well as the (learnable) PE of the defined word. They use BPE subword segmentation for all words; instead of the PE of the defined word, they choose the PE of its first subword. In contrast to Niehues (2021), Zhong and Chiang (2022) find that the model with BPE can inflect dictionary definitions without switching to character-level tokenization.

3 Methodology

Our approach falls squarely into the dictionaries-as-text category: given a source sentence, we retrieve relevant entries from a bilingual dictionary and include them in the source sentence. To decide which entries to include, we use a source-language lemmatizer, which should be more reliable and faster than fuzzy matching. To represent the input, we use *attention masking* instead of positional encodings since we suspect that attention is a more natural mechanism by which an encoder-decoder model can associate definitions (keys) with rare words (queries).

In this section, we break down our approach for using a bilingual dictionary for machine translation with a transformer-based, encoder-decoder model into the following steps: (1) headword selection, (2) definition retrieval, and (3) attention masking.

3.1 Headword Selection

In order to classify a source word as *rare*, we compare the number of occurrences in the training data against a frequency threshold that we choose from a hyperparameter search. For a given source word, we say that a word w is *rare* if (a) it has both a frequency below the threshold and an entry in the dictionary, or (b) if w does not meet either of the above criteria, but its lemmatized form meets both.

3.2 Definition Retrieval

If a rare/unknown word is present in the dictionary, we retrieve its definition(s). Otherwise, we first use a lemmatizer and check if the dictionary contains the lemma for the rare/unknown word. Then, we append the definition(s) to the source sentence following the end-of-sentence token <EOS>.

3.3 Attention Masking

The input now contains a source sentence augmented with dictionary definitions, both segmented into subwords using BPE. To inform the model about the structure of the input, we use attention masking (Shen et al., 2018).

Let n be the input length (source subwords plus definition subwords), and let d be the dimensionality of the model's hidden vectors. In standard attention, we compute, for each head h, a matrix of attention weights $\alpha_h \in \mathbb{R}^{n \times n}$:

$$\alpha_h = \text{softmax}\left(\frac{\mathbf{Q}_h \mathbf{K}_h^T}{\sqrt{d}}\right)$$

where $\mathbf{Q}_h, \mathbf{K}_h \in \mathbb{R}^{n \times d}$ are the query and key matrices, respectively, for head h.

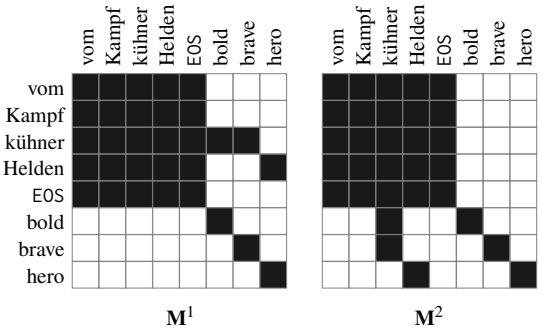

Figure 2: Our system uses two attention masks with learnable strengths. Rows are queries; columns are keys/values. Black = not masked; white = masked. Mask $\mathbf{M}^1$ allows each source word to attend to its definitions (if any). Mask $\mathbf{M}^2$ allows each definition word to attend to the word it defines.

We construct two masks (see Figure 2). Both masks allow all source subwords to attend to all source subwords, and all definition subwords to attend to all subwords in the same definition. Note that *kühner* has two definitions, which cannot attend to each other. Mask $\mathbf{M}^1$ allows each source subword to attend to its definitions (if any). Mask $\mathbf{M}^2$ allows each definition subword to attend to the word it defines. Mathematically, we represent each mask as a matrix $\mathbf{M}^k \in \{0, 1\}^{n \times n}$, where $\mathbf{M}^k_{ij} = 1$ means that subword i cannot attend to subword j.

The attention masks are applied softly, with learnable weights. We combine the masks as follows:

$$\alpha_h = \text{softmax}\left(\frac{\mathbf{Q}_h \mathbf{K}_h^T}{\sqrt{d}} - \sum_{k=1}^{m} \exp\left(s_{k,h}\right) \mathbf{M}^k\right)$$

where $m = 2$ is the number of masks and $s_{k,h} \in \mathbb{R}$ is the learnable strength for mask k and head h. We apply the exponential function component-wise to each $s_{k,h}$ to ensure that every element of the summation is positive. The aggregate attention mask is then subtracted from the standard dot-product attention. In

this way, the model can decide if and/or when the dictionary definitions are useful and adjust the strengths of the attention masks accordingly (McDonald and Chiang, 2021).

4 Experiments

In this section, we describe our translation model, the source-side lemmatizer, and the bilingual dictionary used hereafter throughout the paper.

4.1 Translation Model

To experiment with the internal architecture, we implement an encoder-decoder model from scratch using `PyTorch` (Paszke et al., 2019).[1] For the encoder/decoder, we use a transformer model (Vaswani et al., 2017). Hidden vectors have $d = d_{\text{model}} = 512$ dimensions, and feed-forward networks have $d_{\text{FFN}} = 2048$ dimensions. The encoder and decoder each have 6 layers, each with 8 attention heads. We apply dropout to all embedding, feed-forward, and attention layers with a probability of 0.1. Instead of layer normalization, we use `FixNorm` and `ScaleNorm`, which have been shown to improve translations in the low-resource setting (Nguyen and Salazar, 2019).

All models are trained on NVIDIA A10 GPUs. We use negative log-likelihood for training with a batch size of 4096, a label smoothing value of 0.1, and an initial learning rate of $3 \cdot 10^{-4}$, which we decay by a factor of 0.8 with a patience of 3 and a minimum learning rate of $5 \cdot 10^{-5}$. In addition, we do early stopping if our model trains for 20 epochs without improvement or exceeds a maximum of 250 epochs. Finally, we filter through the training data by removing empty translations, duplicate sentence pairs, sentences longer than a maximum length of 256, and sentence pairs with a source:target length ratio greater than 1.3. We also normalize the punctuation in both the source and target languages.

4.2 Training/Evaluation Data

For German to English translation, we use data from the WMT22 shared task: Europarl v10 for training (Koehn, 2005), newstest2019 for validation, and newstest2022 for testing. For tokenization, we use `sacremoses`,[2] an implementation of Moses (Koehn et al., 2007), at the word-level and `subword-nmt`,[3] an implementation of BPE, at the subword-level. For evaluation, we use a fork of `sacrebleu` (Post, 2018) for BLEU (Papineni et al., 2002) and MacroF1 (Gowda et al., 2021).[4]

The Europarl corpus for German to English has 1,778,520 sentences, with 1,379,973 remaining after cleaning. We apply BPE with 32,000 merge operations and a dropout probability of 0.1 to obtain a shared vocabulary size of 32,469. The newstest2019 validation set and newstest2022 test set contain 2,000 and 1,984 sentences, respectively. To measure translation performance in a low-resource setting, we limit the Europarl corpus to the first 250,000 sentences. The smaller training set has 190,686 sentences remaining after cleaning. We apply BPE with 8,000 merge operations and a dropout probability of 0.1 to obtain a shared vocabulary size of 8,348.

Regarding the difficulty of finding and/or curating extensive dictionaries for low-resource languages, the available Uyghur-English data for the DARPA LORELEI Year 1 evaluation (Hermjakob et al., 2018), for example, consisted of 99k sentences of parallel text and 240k dictionary entries, so there are cases where the amount of dictionary data available is extensive compared to the amount of parallel text available. Given the lack of available training data for low-resource languages, we would argue that hiring linguists to construct bilingual dictionaries offers a greater overall benefit to the community of native speakers and those wishing to document/preserve/revitalize the language than simply hiring translators to expand the available corpora, as the usefulness of dictionaries extends beyond NLP applications (Garrette and Baldridge, 2013).

To evaluate the performance of our model on out-of-domain translation, we combine the Medline test sets from the WMT20 (Bawden et al., 2020), WMT21 (Yeganova et al., 2021), and WMT22 (Neves et al., 2022) biomedical tasks, removing any duplicate sentence pairs. However, the parallel text is misaligned, so we use the provided alignment files to construct the test set, filtering out all sentence pairs not labeled as `OK`. The final test set has 1,073 sentences. Table 3 shows an example sentence from the biomedical test set along with the reference translation.

[1] https://github.com/kennethsible/dictionary-attention

[2] https://github.com/hplt-project/sacremoses

[3] https://github.com/rsennrich/subword-nmt

[4] https://github.com/isi-nlp/sacrebleu

Training Corpus	Model	News		Biomedical	
		BLEU	MacroF1	BLEU	MacroF1
Europarl (Limited)	Baseline	22.1	18.1	18.2	18.5
	Parallel	22.3	18.2	18.2	18.4
	DPE	22.4	18.4	18.3	18.6
	Masking	**23.4**	**20.0**	**19.1**	**19.9**
Europarl (Full)	Baseline	30.4	25.4	23.8	25.8
	Parallel	30.5	25.4	24.0	25.9
	DPE	31.1	26.3	24.3	26.5
	Masking	**31.2**	**26.8**	**24.4**	**26.9**

Table 1: Baseline refers to the translation model without any dictionaries, Parallel includes a bilingual dictionary as parallel text, DPE appends dictionary definitions and uses positional encodings (Zhong and Chiang, 2022), and Masking (ours) appends dictionary definitions and uses attention masking. To construct Europarl (Limited), we only use the first 250,000 sentences (<10%) of the 1.8 million in Europarl (Full).

4.3 Lemmatizer and Dictionary

For the German lemmatizer, we used the spaCy model `de_core_news_sm`[5] with only the tok2vec, tagger, and lemmatizer enabled in the NLP pipeline. For the bilingual dictionary, we used the most recent development version of the German to English bilingual dictionary provided by TU Chemnitz.[6] To prepare the data for our model, we filtered out:

- All dictionary headwords labeled non-alphabetic in Python, excluding hyphenated compound (*e.g.*, `im eigenen Tempo`).
- All dictionary metadata contained in grouping symbols, such as part-of-speech and gender (*e.g.*, masculine noun `{m}`, transitive verb `{vt}`, biological term `[biol.]`, Austrian dialect `[Ös.]`).
- All dictionary abbreviations used for nominative, accusative, dative, and genitive objects (*e.g.*, `jdm.`, `jdn.`, `jds.`, and `etw.`).
- All German prepositional phrases of the form: preposition + abbreviation (*e.g.*, `bei jdm./etw.`).
- The German reflexive pronoun *sich* whenever preceding a headword (*e.g.*, `sich anschließen`).

The German to English dictionary, after cleaning and applying the filters, has 302,061 entries.

[5] https://spacy.io/models/de

[6] https://ftp.tu-chemnitz.de/pub/Local/urz/ding/de-en-devel/

4.4 Experimental Setup

In addition to our model (Masking), we trained three baseline models: a translation model without any dictionaries (Baseline), a model that includes a bilingual dictionary as parallel text (Parallel), and a model that uses dictionary positional encodings (DPE) (Zhong and Chiang, 2022). For DPE and Masking, we append dictionary definitions to source sentences containing rare words. All models were trained on two datasets: Europarl (Limited) and Europarl (Full).

4.5 Hyperparameter Search

By appending dictionary definitions, we introduce two hyperparameters in the model: the frequency threshold for rare words and the number of definitions (or word senses) appended for each rare word. In our experiments, we used frequency thresholds of 5, 10, 15, 25, and 50, and restricted the number of definitions appended to 1, 5, 10, and unbounded.

5 Results

In this section, we report and analyze the results of our experiments described in the previous section. We found that appending definitions for rare words and using attention masking (Masking) improved translation performance over the baseline models: Baseline, Parallel, and DPE.

Furthermore, we observed that using a lower frequency threshold during training and increasing that threshold during inference resulted in the largest performance improvement. We speculate that this

Source	Mit seiner **Tarnkappe** **entkam** **Siegfried** dem **kampferprobten** **Ritter**, einem
	`Mit seiner T@@ ar@@ n@@ k@@ app@@ e ent@@ kam Sie@@ g@@ fri@@ ed dem kampf@@ er@@ prob@@ ten R@@ it@@ ter , einem`
	with his invisibility cloak evaded Sigurd the battle-hardened knight a
	Todfeind, und **schlich** sich aus der Burg.
	`To@@ d@@ fein@@ d , und sch@@ lich sich aus der Burg .`
	deadly enemy and crept himself out of the castle

	Tarnkappe: {invisibility cloak}; **entkam**: {evaded, escaped, got away}; **Siegfried**: {Sigurd}; **kampferprobt**: {battle-seasoned, battle-hardened, battle-tested, combat proven}; **Ritter**: {knight, knights, companion of the order of knighthood, chevalier}; **Todfeind**: {deadly enemy, mortal enemy}; **schlich**: {crept, slunk, tiptoed}
Reference	With his invisibility cloak, Siegfried evaded the battle-hardened knight, a deadly foe, and crept out of the castle.
Baseline	With his cap, Siegfried escaped the tried and tested ritter, a death-enemy, and smashed from the castle.
Parallel	With his cap, Siegfried escaped the tried and tested Ritter, a death enemy, and came out of the castle shamefully.
DPE	With his glasscloak, Sigurd escaped the fighter's knight, a deadly enemy, and crept out of the castle.
Masking	With his invisibility cloak, Sigurd escaped the battle-tested knight, a deadly enemy, and crept out of the castle.
Apple	With his camouflage cap, Siegfried escaped the battle-tested knight, a mortal enemy, and crept out of the castle.

Table 2: On a German sentence (Source), our system's output (Masking) is closer to the Reference than the Baseline system's, even when the dictionary is included in the baseline system's training data (Parallel) or dictionary positional encodings (Zhong and Chiang, 2022) are used instead of attention masking (DPE). Even Apple's Translate app translates *Tarnkappe* over-literally as *camouflage cap*. Rare words are written in boldface. The Reference sentence was written by the first author to demonstrate multiple rare words with a variety of parts of speech and inflections, and a native German speaker translated it into the Source sentence.

Source	Typisch für ein **konjunktivales** **Lymphom** ist eine **lachsfarbene** **Schwellung**. `Typ@@ isch für ein kon@@ jun@@ ktiv@@ ales L@@ ymp@@ ho@@ m ist eine la@@ chs@@ far@@ bene Schwell@@ ung .` typical for a conjunctival lymphoma is a salmon-colored swelling
	Lymphom: {lymphoma}; **lachsfarben**: {salmon, salmon-coloured, salmon-colored}; **Schwellung**: {swelling-up, swelling, puffiness, tumescence, intumescence, intumescentia, tumentia, tumefaction, tumidity, turgescence, turgidity, engorgement}
Reference	A salmon-colored swelling is typical for conjunctival lymphoma.
Baseline	A lax threshold is typical of a lax lymphom in economic terms.
Parallel	A low level threshold is typical of a cyclical lymphom.
DPE	A cyclical lymphom is typically characterised by a lame threshold.
Masking	A salmon-coloured lymphoma is typical of a cyclical lymphoma.
↳ Restricted	A salmon-coloured swelling is typical of a current lymphoma.
↳ Updated	A salmon-coloured swelling is typical of a conjunctival lymphoma.

Table 3: On a German sentence (Source) from the biomedical dataset, our system's output (Masking) is closer to the Reference than the Baseline system's, even when the dictionary is included in the baseline system's training data (Parallel) or dictionary positional encodings (Zhong and Chiang, 2022) are used instead of attention masking (DPE). Rare words are written in boldface. We also edited the input manually for demonstration purposes: For Restricted, the number of definitions appended has been restricted to 3 since *Schwellung* has 12, which causes the model to struggle. For Updated, we restricted the number of definitions appended to 3 and added a definition for *konjunktival* 'conjunctival' to the dictionary (not previously present).

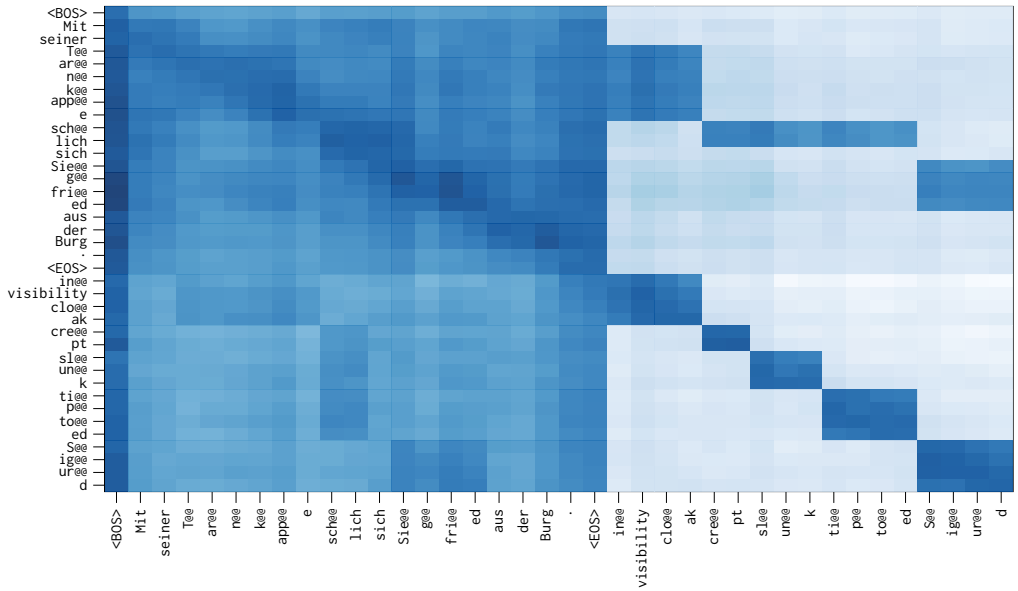

Figure 3: The attention scores of the Masking model for the German sentence: "Mit seiner Tarnkappe schlich sich Siegfried aus der Burg" with the definition string "invisibility cloak crept slunk tiptoed Sigurd." The attention scores are summed for all encoder layers and attention heads. We observe both attention masks being utilized by the model.

behavior is a result of larger thresholds incorrectly classifying unknown compound words (*i.e.*, those not occurring in the training data) as rare when they are already translatable by the baseline models as a result of subword tokenization. If we are attempting to teach the model to use definitions, including them when they are not necessary may actively work against our training objective.

During a hyperparameter search, we found that using 10 for the frequency threshold and 10 for the maximum number of definitions appended yields the largest improvement in translation performance. Table 1 compares our model against the three baseline models for both the general (news) and out-of-domain (biomedical) datasets. All metrics reported in the table were averaged over 5 random restarts and statistical significance was verified with paired t-tests. Masking (our model) outperforms the three baseline models on both metrics and all improvements are statistically significant (p-value < 0.05), except for the BLEU improvement over DPE for Europarl (Full).

In the low-resource setting, DPE struggles to improve over the baseline models, while Masking has the largest BLEU improvement, demonstrating a boost in low-resource translation performance. In the high-resource setting, although DPE and Masking are not significantly different in BLEU, they are significantly different in MacroF1. Since rare words are less frequent in the high-resource setting, the BLEU improvement of DPE and Masking over Baseline and Parallel is not as large. However, Masking has the largest MacroF1 improvement, demonstrating a boost in rare word translation performance.

In Table 2, we compare candidate translations of a German sentence containing rare words against an English reference. The sentence was written in English, and translated by a native German speaker, to demonstrate the capability and robustness of our model in using the dictionary. The German sentence contains seven rare words of varying part-of-speech, including adjective declension and verb conjugation. In the Source row, the English glosses are shown beneath each German word to match the Reference translation along with the corresponding subword tokenization. To reduce the sentence length, the definitions are listed separately instead of appended to the German sentence.

Baseline and Parallel contain several incorrect

translations of rare words. In particular, we observe that *Tarnkappe* and *Todfeind* were translated over-literally, with the first noun in the compound being dropped all together. Even the Apple Translate app translated *Tarnkappe* over-literally as *camouflage cap*. The DPE model, instead of dropping the first noun like Baseline/Parallel, used a seemingly random noun and translated the second over-literally. Only the Masking model correctly translated *Tarnkappe* as *invisibility cloak*. In fact, Masking used at least one definition for every rare word, getting the closest to the Reference.

In Table 3, we compare candidate translations of a German sentence taken directly from the Medline test set. The sentence contains four rare words, but our dictionary has no definition for *konjunktival*. *Lymphom*, despite having a definition, is copied to English sentence by Baseline, Parallel, and DPE. Masking correctly translates *Lymphom* and *lachsfarben*, but all models mistranslate *Schwellung*. We found that Masking often ignores definitions if there are too many appended for a given rare word. To demonstrate, we restricted the number of definitions for *Schwellung* to 3 and see that the model correctly translates the word. We also succeeded in translating *konjunktival* correctly by adding the English definition to the dictionary, demonstrating that the dictionary coverage is a limiting factor.

In Figure 3, we use an attention heat map to visualize the attention scores for a German sentence. The sentence shown is a trimmed version of the example in Table 2. To build the heat map, we summed the attention scores for every encoder layer and every attention head. We see that the attention masks shown in Figure 2 are clearly visible in the heat map. However, the model decided to put more emphasis on the first mask than the second, which is done by adjusting the mask strengths.

6 Discussion

Rare Word Classification As mentioned previously, compound words that do not occur in the training data may still be accurately translated as a result of subword segmentation, suggesting that frequency is not an ideal or reliable metric for classifying rare words. In the future, frequency could be replaced with a source-side estimation of model confidence in the translation of rare words.

Incorrect Lemmatization We could not find an acceptable lemmatizer for the German language since even spaCy would occasionally misidentify the lemma for, *e.g.*, a declined adjective or a past participle. Furthermore, no lemmatizer that we found could correctly identify the infinitive form for separable verbs or *trennbare Verben*, a common class of verbs in the German language. In the future, we could explore more robust lemmatization techniques or the inclusion of inflected forms in the dictionary.

Lemmatization Ambiguities We have identified several cases where lemmatization causes the model to use a definition that is not grammatically correct in the context of the source sentence. For example, if the past tense form of a verb is not present in the dictionary and the definition for the infinitive form is used, the model often avoids inflecting the infinitive form to the correct tense unless the sentence contains, *e.g.*, an auxiliary verb. Similarly, nouns ending in –er in German have no plural ending, which creates an ambiguity as to whether the English definition should be plural. A dictionary that directly contains inflected forms may resolve such ambiguities.

Definition/Word Sense Pruning We appended definitions for each word sense and part-of-speech with the assumption that the model could learn to leverage syntactic or semantic knowledge of the source sentence to select an appropriate translation for the rare words from among those definitions appended. However, we find that the model is often spoiled for choice, in that the model may use an inappropriate definition, or none at all, if there are too many without a clear way to disambiguate. In the future, we could implement a strategy to select the most relevant definitions or limit the number of appended definitions per rare word, such as pruning based on document-level context or prior domain knowledge.

Phrases and Compound Words We appended definitions only for single words, which includes both hyphenated and concatenated compound words in German, but did not consider phrases whose translations may not be directly deducible from the constituent words. Similarly, we did not consider separating compound words into the constituent words and recursively searching for definitions if the compound words are not present in the dictionary themselves since subword segmentation often handles these.

7 Conclusion

In this paper, we proposed using bilingual dictionaries and attention masking to improve translation performance for rare words, a problem that encoder-decoder models continue to struggle with in MT. Our method was to append definitions to source sentences for low-frequency words and use attention masking to associate rare words with their definitions. We found that our method improved MT performance by up to 1.0 BLEU and 1.6 MacroF1. In the future, we are interested in incorporating other external knowledge sources, such as monolingual dictionaries and knowledge graphs, to reduce translation ambiguity and further improve the translation of rare words.

8 Limitations

The following are two limiting factors of our masking approach to including bilingual dictionaries in machine translation: (1) the quality and coverage of the lemmatizer and/or dictionary is a bottleneck to further improvement and (2) appending definitions increases sentence length and therefore runtime.

9 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-2137396.

References

Arthur, P., Neubig, G., and Nakamura, S. (2016). Incorporating discrete translation lexicons into neural machine translation. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas. Association for Computational Linguistics.

Bawden, R., Di Nunzio, G. M., Grozea, C., Jauregi Unanue, I., Jimeno Yepes, A., Mah, N., Martinez, D., Névéol, A., Neves, M., Oronoz, M., Perez-de Viñaspre, O., Piccardi, M., Roller, R., Siu, A., Thomas, P., Vezzani, F., Vicente Navarro, M., Wiemann, D., and Yeganova, L. (2020). Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Graham, Y., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., and Negri, M., editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 660–687, Online. Association for Computational Linguistics.

Chatterjee, R., Negri, M., Turchi, M., Federico, M., Specia, L., and Blain, F. (2017). Guiding neural machine translation decoding with external knowledge. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., and Kreutzer, J., editors, *Proceedings of the Second Conference on Machine Translation*, pages 157–168, Copenhagen, Denmark. Association for Computational Linguistics.

Dinu, G., Mathur, P., Federico, M., and Al-Onaizan, Y. (2019). Training neural machine translation to apply terminology constraints. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data augmentation for low-resource neural machine translation. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Garrette, D. and Baldridge, J. (2013). Learning a part-of-speech tagger from two hours of annotation. In Vanderwende, L., Daumé III, H., and Kirchhoff, K., editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147, Atlanta, Georgia. Association for Computational Linguistics.

Gowda, T., You, W., Lignos, C., and May, J. (2021). Macro-average: Rare types are important too. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1138–1157, Online. Association for Computational Linguistics.

Hasler, E., de Gispert, A., Iglesias, G., and Byrne, B. (2018). Neural machine translation decoding with terminology

constraints. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana. Association for Computational Linguistics.

Hermjakob, U., Li, Q., Marcu, D., May, J., Mielke, S. J., Pourdamghani, N., Pust, M., Shi, X., Knight, K., Levinboim, T., Murray, K., Chiang, D., Zhang, B., Pan, X., Lu, D., Lin, Y., and Ji, H. (2018). Incident-Driven Machine Translation and Name Tagging for Low-resource Languages. *Machine Translation*, 32(1):59–89.

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Koehn, P., Marie, B., Monz, C., Morishita, M., Murray, K., Nagata, M., Nakazawa, T., Popel, M., Popović, M., and Shmatova, M. (2023). Findings of the 2023 Conference on Machine Translation (WMT23): LLMs are here but not quite there yet. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In Ananiadou, S., editor, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

McDonald, C. and Chiang, D. (2021). Syntax-based attention masking for neural machine translation. In Durmus, E., Gupta, V., Liu, N., Peng, N., and Su, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 47–52, Online. Association for Computational Linguistics.

Minh-Cong, N.-H., Ngo, V. T., and Nguyen, V. V. (2022). A simple and fast strategy for handling rare words in neural machine translation. In Hanqi, Y., Zonghan, Y., Ruder, S., and Xiaojun, W., editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 40–46, Online. Association for Computational Linguistics.

Neves, M., Jimeno Yepes, A., Siu, A., Roller, R., Thomas, P., Vicente Navarro, M., Yeganova, L., Wiemann, D., Di Nunzio, G. M., Vezzani, F., Gerardin, C., Bawden, R., Estrada, D. J., Lima-Lopez, S., Farre-Maduel, E., Krallinger, M., Grozea, C., and Neveol, A. (2022). Findings of the WMT 2022 biomedical translation shared task: Monolingual clinical case reports. In *Proceedings of the Seventh Conference on Machine Translation*, pages 694–723, Abu Dhabi. Association for Computational Linguistics.

Nguyen, T. Q. and Salazar, J. (2019). Transformers without tears: Improving the normalization of self-attention. In Niehues, J., Cattoni, R., Stüker, S., Negri, M., Turchi, M., Ha, T.-L., Salesky, E., Sanabria, R., Barrault, L., Specia, L., and Federico, M., editors, *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Niehues, J. (2021). Continuous learning in neural machine translation using bilingual dictionaries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 830–840, Online. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, USA. Association for Computational Linguistics.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, number 721, pages 8026–8037. Curran Associates Inc., Red Hook, NY, USA.

Post, M. (2018). A call for clarity in reporting BLEU scores. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Post, M. and Vilar, D. (2018). Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Shen, T., Jiang, J., Zhou, T., Pan, S., Long, G., and Zhang, C. (2018). DiSAN: directional self-attention network for RNN/CNN-free language understanding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Tan, L., van Genabith, J., and Bond, F. (2015). Passive and pervasive use of bilingual dictionary in statistical machine translation. In Babych, B., Eberle, K., Lambert, P., Rapp, R., Banchs, R. E., and Costa-jussà, M. R., editors, *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 30–34, Beijing. Association for Computational Linguistics.

Thompson, B., Knowles, R., Zhang, X., Khayrallah, H., Duh, K., and Koehn, P. (2019). HABLex: Human annotated bilingual lexicons for experiments in machine translation. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1382–1387, Hong Kong, China. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wu, Q., Xing, C., Li, Y., Ke, G., He, D., and Liu, T.-Y. (2021). Taking notes on the fly helps language pre-training. In *International Conference on Learning Representations*.

Xu, H., Kim, Y. J., Sharaf, A., and Awadalla, H. H. (2024). A paradigm shift in machine translation: Boosting translation performance of large language models. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Yeganova, L., Wiemann, D., Neves, M., Vezzani, F., Siu, A., Jauregi Unanue, I., Oronoz, M., Mah, N., Névéol, A., Martinez, D., Bawden, R., Di Nunzio, G. M., Roller, R., Thomas, P., Grozea, C., Perez-de Viñaspre, O., Vicente Navarro, M., and Jimeno Yepes, A. (2021). Findings of the WMT 2021 biomedical translation shared task: Summaries of animal experiments as new test set. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., and Monz, C., editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 664–683, Online. Association for Computational Linguistics.

Zhang, J. and Zong, C. (2016). Bridging neural machine translation and bilingual dictionaries.

Zhang, T., Zhang, L., Ye, W., Li, B., Sun, J., Zhu, X., Zhao, W., and Zhang, S. (2021). Point, disambiguate and copy: Incorporating bilingual dictionaries for neural machine translation. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3970–3979, Online. Association for Computational Linguistics.

Zhong, X. J. and Chiang, D. (2022). Look it up: Bilingual dictionaries improve neural machine translation. arXiv:2010.05997.