Improving Machine Translation of Rare and Unseen Word Senses

[1]Viktor Hangya, [2]Qianchu Liu, [1]Dario Stojanovski,
[1]Alexander Fraser and [2]Anna Korhonen
[1]Center for Information and Language Processing, LMU Munich
{hangyav,stojanovski,fraser}@cis.lmu.de
[2]Language Technology Lab, TAL, University of Cambridge, UK
{ql261,alk23}@cam.ac.uk

Abstract

The performance of NMT systems has improved drastically in the past few years but the translation of multi-sense words still poses a challenge. Since word senses are not represented uniformly in the parallel corpora used for training, there is an excessive use of the most frequent sense in MT output. In this work, we propose CMBT (**C**ontextually-**m**ined **B**ack-**T**ranslation), an approach for improving multi-sense word translation leveraging pre-trained cross-lingual contextual word representations (CCWRs). Because of their contextual sensitivity and their large pre-training data, CCWRs can easily capture word senses that are missing or very rare in parallel corpora used to train MT. Specifically, CMBT applies bilingual lexicon induction on CCWRs to mine sense-specific target sentences from a monolingual dataset, and then back-translates these sentences to generate a pseudo parallel corpus as additional training data for an MT system. We test the translation quality of ambiguous words on the MuCoW test suite, which was built to test the word sense disambiguation effectiveness of MT systems. We show that our system improves on the translation of difficult unseen and low frequency word senses.

1 Introduction

Recent NMT systems have remarkable performance for many languages (Vaswani et al., 2017; Xia et al., 2019) but there are still numerous areas for improvement. One such important area concerns the disambiguation and translation of multi-sense words. It is particularly challenging to MT systems as sense distribution is skewed with some senses rarely seen or missing in the parallel corpora. This results in the MT system producing translation errors for these rare/unseen senses, causing incomprehensible output sentences.

In this work we aim at improving the translation of rare and unseen senses of ambiguous words.

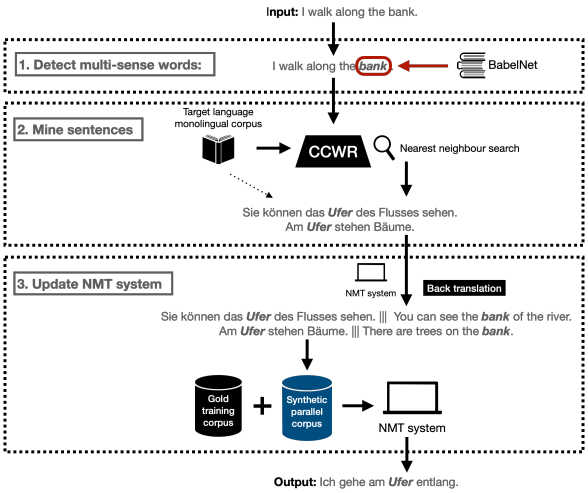

Figure 1: The pipeline of CMBT. Note that each step of the pipeline is run on the full corpus (source side of MuCoW test set). Here we just show the procedure on a single sentence as an illustration.

Previously Tang et al. (2018, 2020) showed that encoder-decoder based NMT systems integrate information relevant for WSD into the encoder hidden states, but finding the correct sense is still a challenging task and NMT systems are biased toward the most frequent senses of words (Liu et al., 2018). Disambiguation errors are often due to over-reliance on training data artifacts, such as frequent word co-occurrences (e.g. *hot spring* is always translated as the thermal activity and not as a season), instead of a deeper understanding of the multi-sense words given the input sentences (Emelin et al., 2020). Additionally, MT systems tend to learn and use frequent words more often and disregard less frequent ones (Vanmassenhove et al., 2019). Previous work has improved the translation of ambiguous words, e.g., by leveraging lexical resources (Pu et al., 2018) or sense specific embeddings (Liu et al., 2018), but they are restricted to the senses seen in the parallel training corpus and not trained on missing senses.

In contrast, we propose a method to mine addi-

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 614–624
November 10–11, 2021.

tional data containing the contextual translations of rare and unseen senses without relying on them being in parallel corpora. Our method, called CMBT (**C**ontextually-**m**ined **B**ack-**T**ranslation), relies on contextualized cross-lingual word representations (CCWRs) to translate source language multi-sense words and find target language sentences containing the translations of the right senses. We then build a synthetic parallel corpus by back-translating these sentences, making sure that the original multi-sense words are contained on the source side, in order to extend the training corpus for better sense coverage. We illustrate our method in Figure 1. The advantage of our approach is that CCWRs, such as mBERT (Devlin et al., 2017) or XLM-R (Conneau et al., 2020), can be trained on cheap and large monolingual corpora covering a wide frequency range of word senses. By leveraging CCWR-mined sentences containing the translations of these senses in the form of a synthetic parallel corpus, our MT system is not restricted to frequent senses seen in the parallel training corpus.

We test our approach on English→German using the *MuCoW* test suite (Raganato et al., 2019, 2020). Although it was built to test overall WSD performance of MT systems, we create subsets of the provided training corpus to test on unseen and rare senses more directly. Our experiments show that using mined sentences as additional data for our NMT systems consistently improves the translation performance (F_1) of rare and unseen senses. Our proposed approach can be effectively applied to other language pairs as well, since the required resources are widely available.

2 Related Work

The problem of WSD is long-standing and extensively studied. Multiple neural systems were proposed, e.g., by relying on sequence-to-sequence architectures (Raganato et al., 2017), using sense embeddings (Kumar et al., 2019) or pre-trained language models (Pasini et al., 2021). It was shown that WSD positively impacts the performance of downstream applications, such as information retrieval (Zhong and Ng, 2012), sentiment analysis (Pilehvar et al., 2017) or topic classification (Shimura et al., 2019).

WSD is an important problem for MT as well. Previously it was shown that the translation performance can be improved by integrating word sense information into MT systems. In (Pu et al., 2017) sense labels were assigned to each multi-sense word using K-means clustering, which served as additional information for a statistical MT system. Similarly, explicit word sense information using WordNet was integrated into NMT systems in (Pu et al., 2018). Liu et al. (2018) leveraged sense embeddings induced by specialized LSTM modules, while lexical chains of semantically similar words within a document were employed in (Rios et al., 2017). Although these approaches do improve WSD performance, they are restricted to the senses seen frequently in the parallel training corpus.

In contrast, we focus on the improvement of senses that are missing or very rare by building a synthetic parallel corpus containing these senses using back-translation (Sennrich et al., 2016). Similarly, Huck et al. (2019) back-translated a carefully selected set of sentences to improve the translation of out-of-vocabulary (OOV) words, i.e., words that are contained in the text to be translated but not in the training corpus. They used bilingual fastText (Bojanowski et al., 2017) embeddings to find all translations of OOVs independent of their contexts. In contrast, we consider the whole sentence when translating multi-sense words in order to determine the right translation of the right sense used in the right context using CCWRs, which we show to be crucial to improve missing and rare sense translation. Arthaud et al. (2021) proposed a data augmentation approach to adapt MT systems to novel vocabulary in human-submitted translations using CCWRs. They generate training samples for the novel words by mining parallel sentence pairs with similar contexts and adding the novel words to them. In contrast, our approach does not rely on parallel sentences, using only monolingual data and back-translation.

Various datasets were proposed to test WSD, such as those released by the series of Senseval (Edmonds and Cotton, 2001; Mihalcea et al., 2004) and SemEval (Agirre et al., 2010; Navigli et al., 2013; Moro and Navigli, 2015) shared tasks. While most datasets are monolingual, Pasini et al. (2021) introduced XL-WSD supporting 18 languages allowing to evaluate zero-shot cross-lingual WSD approaches. To test how well MT systems can disambiguate multi-sense words in their outputs Rios et al. (2017) created the parallel corpus called ContraWSD where for each source sentence containing an ambiguous word two translations are given with

the correct and incorrect senses respectively which have to be scored by the MT systems. The MuCoW dataset was introduced for a more direct evaluation where instead of scoring given target language sentences the translations of multi-sense words in the MT systems' outputs are evaluated (Raganato et al., 2019, 2020). We use MuCoW to evaluate our approach.

3 Approach

The goal of CMBT is to incorporate context-dependent word translation that is able to deal with rare and unseen senses and leverage cheap monolingual data as additional training data for our NMT system for better multi-sense word translation. The main steps of our approach are the following: i) we detect multi-sense words in the source side of test corpus using BabelNet (Navigli and Ponzetto, 2012), which ii) we translate using CCWRs and mine target language sentences containing these translations. iii) We back-translate these sentences to the source language using a baseline NMT system. We use a special marker placed in the target language sentences, which are replaced with the multi-sense words on the source side, in order to ensure the presence of rare and unseen senses in the new corpus. Finally, we fine-tune our base NMT system using the gold and additional synthetic parallel data. We summarize the pipeline in our approach in Figure 1 and detail the three main steps below:

3.1 Multi-Sense Word Detection

As the first step, we identify multi-sense words in the test corpus relying on BabelNet, a publicly available multilingual lexical resource covering 284 languages (Navigli and Ponzetto, 2012). Since the MuCoW dataset focuses on nouns only, we first take all English nouns from the test corpus.[1] We then filter out single sense nouns by keeping only those which are contained in at least two synsets. However BabelNet has a very fine grained set of synsets which would result in a list containing many single sense nouns as well due to their inclusion in multiple synsets. Thus before filtering we merge some of the synsets using English-German interlingual links in BabelNet which specify possible German translations of the words in a given English synset. More precisely, we merge English synsets which have overlapping sets of translations. The filtering procedure using the merged synsets resulted in 3 732 multi-sense nouns containing 181 out of the 206 gold multi-sense words in the MuCoW test corpus.[2]

We note that although BabelNet covers a large set of languages, the language of the application area might not be supported. However, CMBT only requires a list of source language multi-sense words as input which can be acquired using unsupervised WSD systems as well, such as the word embeddings based *SenseGram* (Pelevina et al., 2016). We argue that our approach is robust against false positive multi-sense words, since we would mine sentences containing their single sense, thus the use of a high recall list is preferable in such cases.

3.2 Sentence mining

Given the multi-sense words we mine target language sentences containing the translations of their different senses. However we do not mine all possible senses of the words but only those which are contained in the input corpus to be translated, i.e., the source side of the test corpus in our case. For this we perform bilingual token-level sense retrieval (BTSR) (Liu et al., 2019) where the task given a source word in a context (sentence) is to retrieve its translation having the same sense along with a matching target context. More formally, given a $(w_s, c_s) \in V_s \times D_s$ pair the task is to retrieve $(w_t, c_t) \in V_t \times D_t$, such that w_t is the translation of w_s and the sense of w_s in context c_s matches the sense of w_t in c_t. V_s, V_t and D_s, D_t are the vocabularies and the monolingual datasets of the source and target languages respectively.

To mine relevant sentences, we take each multi-sense word contained in each source sentence as the input (w_s, c_s) pairs. Since a given word type is contained in multiple sentences, we perform mining using these sentences individually. As the translation target candidates, we take a target language monolingual corpus (see Section 4.3 for more details) and consider each word in each sentence as a candidate (w_t, c_t) pair. For each source input pair, we take the top-5[3] most similar target pair scored

[1] We used UDPipe (Straka and Straková, 2017) for POS tagging.

[2] Note that BabelNet was also used to build the list of multi-sense words in MuCoW but its output was further refined with parallel data and gold WSD annotations.

[3] Top-5 is common for bilingual lexicon induction.

by:

$$sim((w_s, c_s), (w_t, c_t)) = cos(E_{(w_s,c_s)}, E_{(w_t,c_t)}) \quad (1)$$

where $E_{(w,c)}$ is the CCWR of word w in context c and cos is the cosine similarity of two embeddings. As CCWR of a given word we averaged the corresponding vectors of the upper XLM-R layers (12-24), motivated by the findings of Ethayarajh (2019). We discuss further details of the used CCWR models in Section 4.1. Finally, the retrieved sentences are considered as the output of the mining process and used in the next step.

3.3 NMT System Update

In the last step, we update our baseline English→German NMT system with the gold parallel data and the sentences mined above. Using a system similar to the baseline but built in the reverse direction we back-translate the mined target language sentences by making sure that the original multi-sense word is contained in the back-translation. To achieve this we replace the related word in the target sentence with a special marker which is copied to the source side during translation. After translation we replace the special markers on both sides with the correct words. The following example depicts the process using a mined sentence for the multi-sense word *bank*:

Input:	*Ich gehe am **Ufer** entlang.*
Replace:	*Ich gehe am **<MARK>** entlang.*
Translate:	*I walk along the **<MARK>**.*
Restore:	*I walk along the **bank**.*

To learn the copy mechanism of the special marker we use parallel sentences containing the marker to train the NMT system used for back-translation. More precisely, 1% of the parallel sentences have one randomly selected source word and it's corresponding translation (determined by aligning the parallel data) replaced with the marker. Note that there is a small chance that the MT system does not generate the marker in the output in which case no replacement is performed. Finally, we update our baseline English→German MT system by running further training steps on the concatenated gold and synthetic parallel corpora. For further parameters we refer to Section 4.4.

4 Experimental Setup

4.1 Cross-Lingual Word Representations

As CCWRs we make use of XLM-R large[4] (Conneau et al., 2020), as previous works have shown good context-dependent cross-lingual correspondence in such multilingual models (Ethayarajh, 2019; Liu et al., 2019; Cao et al., 2019). Although they are multi-lingual, it was shown that their cross-lingual performance can be improved by applying an additional mapping step. Thus following Liu et al. (2019), we train a linear orthogonal mapping on XLM-R's context-average word type representations of word pairs extracted from the automatic word alignments in the parallel corpus which is used for MT training as well. The context-average representations are first length normalized and then mean centered prior to alignment as it was shown to improve the mapping quality (Artetxe et al., 2018). On top of the orthogonal mapping, we also apply the meeting-in-the-middle technique proposed by Doval et al. (2020) that learns additional linear mappings of both source and target languages to further improve their alignment. For exact details about the complete mapping process we refer to (Liu et al., 2019).

4.2 Baselines

Other than comparing CMBT with XLM-R to the *baseline* NMT system, we compare the approach to Huck et al. (2019), since it is able to leverage monolingual data to improve the translation of a list of words. More precisely, we translate multi-sense words with BWEs instead of XLM-R, to show the importance of context based word translation for the translation of multi-sense words. Since BWEs tend to rank the translations of words according to their frequency, this approach is comparable to the general back-translation approach, i.e., updating NMT systems on randomly sampled sentences, but focusing more on the ambiguous words. We build 300 dimensional *fastText* skipgram embeddings (Bojanowski et al., 2017) on Wikipedia dumps and align them using the same approach as for XLM-R (Liu et al., 2019). Similarly to CMBT, words are translated using cosine similarity taking top-5 most similar candidates. However, since BWE based bilingual lexicon induction (BLI) is

[4]Besides XLM-R, we experimented with mBERT (Devlin et al., 2017) as well but chose the former due to its superior word retrieval performance (See the experiment's results in Appendix A).

context independent, we pick all sentences containing any of the translations, which are then down sampled to match the number of sentences mined by CMBT for comparability. The rest of the steps, i.e., back-translation and system training, are the same as for CMBT. Our experiments show that although top-5 translations based on BWEs can cover multiple senses of some words, it is important to take the contexts into consideration as well in order to perform i) a more sense specific word translation and ii) mine sentences which do not only contain the target words but have similar contexts compared to the source sentences in the test corpus for an efficient MT tuning.

4.3 Monolingual Dataset

We use $2M$ randomly sampled German Wikipedia sentences for the mining process and restrict the vocabulary for translation candidates to the $500K$[5] most frequent words. We mine relevant sentences for all senses of the detected multi-sense words, including their frequent senses, since the frequency of senses in the training corpus is not known. The mined corpus contains 252 898 unique sentences.

4.4 MT Systems

We train base Transformer NMT models (Vaswani et al., 2017) on the gold parallel data discussed below with early-stopping based on validation perplexity. The model is trained on 4 Nvidia GTX 1080ti GPUs with a per-GPU batch size of 4096 tokens and by delaying stochastic gradient descent updates with a factor of 2. The final model is an average of the best 10 checkpoints, where checkpoints are saved every 500 updates. We use dropout and label smoothing with a value of 0.1.

Fine-tuning this initial system with the concatenation of the gold and synthetic data is done using the same hyper-parameters with early-stopping based on validation perplexity. The average of the best 10 checkpoints is chosen as the initial starting point for fine-tuning.

As a development set we merge newstest 2017-2019. We use a beam size of 4 for back-translation and 5 for the translation of MuCoW. All models are built using fairseq (Ott et al., 2019). The datasets are tokenized using Moses[6]. We use BPE splitting[7] with 32K merge operations computed jointly on the source and target data. Word alignment is performed using fastalign (Dyer et al., 2013).

4.5 MuCoW Dataset

We run experiments on the English→German translation direction of the MuCoW dataset (Raganato et al., 2020). It was created specifically to test translation quality of ambiguous words by specifying the word and its gold sense for each test sentence. The dataset provides small and big training parallel corpora containing $1.2M$ and $3.0M$ sentence pairs respectively. In order to test on rare and unseen senses more directly, we create two subsets of the latter. We remove sentence pairs containing the rarest sense of any of the given multi-sense words to test on *unseen* senses ($2.9M$ pairs). Secondly, we take a random 10% sample of the sentence pairs containing a multi-sense word and all pairs containing no multi-sense words in *sample-10* ($2.4M$ pairs) to test on rare senses. By sampling data uniformly random in case of the latter, we make sure that only the frequency of the multi-sense words gets decreased, while keeping their original sense distribution. Note that we only change the training set to have more word types with unseen and rare senses during training but keep the test set unchanged. Our *baseline* NMT system is trained on these training sets without the additional mined data.

We evaluate our MT systems on word level (F_1) using the official MuCoW evaluation script which calculates precision and recall values as:

$$P = \frac{|correct\ senses|}{|correct\ senses| + |incorrect\ senses|} \quad (2)$$

$$R = \frac{|correct\ senses|}{|test\ cases|} \quad (3)$$

where a test case is an occurrence of a multi-sense word in a test sentence. The dataset provides multiple correct translation options for a given sense, thus an occurrence of a multi-sense word (sense) is correctly translated if any of the translations of the correct sense are contained in the output sentence. A sense is incorrect if any of the translations of the wrong senses of the given multi-sense word are contained in the output sentence. A sense is

[5] We increase the $200K$ limit used in most BLI works in order to cover more rare words.

[6] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

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

train	bin	#	system	$acc@1$	$acc@5$	F_1
unseen	0-0	10.5	baseline	-	-	17.14
			BWEs	6.69	13.51	25.39
			CMBT	**21.88**	**35.32**	**34.80**$\uparrow^{17.66}$
sample-10	0-20	5.9	baseline	-	-	35.53
			BWEs	1.93	4.18	37.70
			CMBT	**15.55**	**26.34**	**47.02**$\uparrow^{11.49}$
	20-40	3.2	baseline	-	-	60.98
			BWEs	7.92	19.78	60.80
			CMBT	**22.29**	**37.34**	**64.49**$\uparrow^{3.51}$

Table 1: Intrinsic and extrinsic evaluation in terms of $acc@n$ and MuCoW F_1 scores. The rare senses in *sample-10* are shown broken down by relative frequency bins, while we present results of missing senses in *unseen*. The number of test cases in thousands per bin is shown in the third column (#). We compare the baseline and the improved MT systems with both BWEs and CMBT. We indicate the improvements (↑) compared to the baseline.

neither correct nor incorrect if none of the possible translations of the multi-sense word is contained. Furthermore sentences are lemmatized, thus all morphological variants of a word are accepted.

We also calculate BLEU scores using *sacrebleu* (Post, 2019) to show general MT performance. In addition, we evaluate the word translation accuracy of XLM-R and BWEs on the gold MuCoW multi-sense words contained in each test sentence. Similarly to BLI (Vulić and Korhonen, 2016), we calculate $acc@n$ $(n \in 1, 5)$ scores by testing if any of the correct translations of the gold sense in a given test example is among the n most similar translation candidates.

5 Results and Discussion

Unseen and rare sense translation We present both the intrinsic performance of BWEs (Huck et al., 2019) or CMBT (XLM-R) based word translation ($acc@n$) and extrinsic MT system based translation (F_1) in Table 1. We show results on senses in the test corpus which are missing from the *unseen* training corpus and detailed results on word senses that are rare (relative frequency compared to the other senses of a given word is between 20% and 40%) and very rare (with relative frequency between 0% and 20%) on *sample-10*.

In terms of $acc@n$ CMBT word translation performs significantly better than BWEs. This is not surprising, since the context independent BWEs predict the same translations for a given multi-sense word for each sentence it is contained in. In contrast, XLM-R shows a better WSD performance

train	system	$acc@1$	$acc@5$	F_1
unseen	baseline	-	-	70.70
	BWEs	17.94	28.74	71.66
	CMBT	**28.30**	**43.90**	**73.51**$\uparrow^{2.81}$
sample-10	baseline	-	-	74.58
	BWEs	17.94	28.74	73.75
	CMBT	**28.30**	**43.90**	**75.86**$\uparrow^{1.28}$

Table 2: Evaluation of all (including frequent) senses when using *unseen* or *sample-10* training sets. Number of overall test cases are $25.3K$ in both sets. BLI results are the same for both training sets as they only affect the NMT system.

by relying on the context in the sentences. Our improved NMT system using CMBT outperforms the baseline system in all setups in terms of F_1. It is especially effective on the unseen and very rare senses due to the additional synthetic sentence pairs containing these senses and their translations. In addition, it is also effective for the rare senses in the higher relative frequency range bin. BWEs based mining is also helpful for the unseen and very rare senses but it is less effective compared to CMBT. Although BWEs are context independent, by taking top-5 translations some of the senses can still be improved. On the other hand, BWEs have minor negative effects for the higher frequency range.

All sense translation We show results on all senses, i.e., senses with relative frequency higher than 40% as well, using the two training sets in Table 2. CMBT is also effective when evaluating on the whole MuCoW test dataset, but its performance is more prominent on the lower frequency ranges. In contrast, BWEs achieved only a slight improvement on *unseen* and some performance drop on *sample-10*.

Lexicon-regularized translation As mentioned, we built the list of English multi-sense words using BabelNet. Since it also contains translation options for each word, we investigate whether we can make use of this additional information. Fortunately, CMBT can be naturally extended to leverage such lexical resources.[8]

During sentence mining with the lexicon-regularized version of our approach (CMBT+), we restrict the set of translation candidates when translating a given word with XLM-R to its possible

[8]In comparison, it is not straightforward how we can add this information into a baseline NMT system where we cannot easily track translation for specific source words.

train	bin	system	F_1
unseen	0-0	ALL+	$25.21\downarrow^{9.59}$
		CMBT+	$34.55\downarrow^{0.25}$
	all	ALL+	$72.00\downarrow^{1.51}$
		CMBT+	$73.83\uparrow^{0.32}$
sample-10	0-20	ALL+	$43.37\downarrow^{3.65}$
		CMBT+	$46.75\downarrow^{0.27}$
	20-40	ALL+	$65.04\uparrow^{0.55}$
		CMBT+	$66.82\uparrow^{2.23}$
	all	ALL+	$75.95\uparrow^{0.09}$
		CMBT+	$76.50\uparrow^{0.64}$

Table 3: Evaluation of the senses per frequency bins as well as all senses using the lexicon-regularized systems on the two datasets. Differences compared to the best system (CMBT in tables 1 and 2) are indicated.

train	bin	baseline	BWE	CMBT
unseen	0-0	23.0	23.2	**23.3**
	all	25.5	25.6	**25.7**
sample-10	0-20	22.3	22.3	**22.6**
	20-40	24.5	24.6	**24.7**
	all	25.0	25.0	**25.1**
big	all	**26.5**	**26.5**	26.4

Table 4: Machine translation performance (BLEU) on the complete MuCoW test set using the unmodified *big* and our two custom training sets. We show results on the missing (0-0) and rare senses (0-20 and 20-40) as well as on the complete test set (all). BLEU score achieved by Raganato et al. (2020) on big-all is 22.6.

translations as given by BabelNet. Furthermore, we mine sentences based on all possible translations (ALL+) instead of taking top-5 ranked by XLM-R. Table 3 shows that the regularized systems achieved improvements compared to CMBT only in the higher frequency ranges but not in the missing or very rare sets. ALL+ achieved only minor improvements overall (*all*) on *sample-10*, while the performance decreased on *unseen*. This indicates that without focused data mining, sentences containing rare and unseen senses are suppressed by the frequent senses. In contrast, CMBT+ achieved improvements on both setups by following the sense distribution of the test set. However, the improvements are marginal which shows that the unregularized CMBT system is already able to retrieve the relevant senses of words without the additional information coming from BabelNet.

BLEU evaluation Finally, we show general MT performance on our training setups including the original MuCoW *big* setup as well for comparison in Table 4. It can be seen that our approach achieved improvements in terms of BLEU as well, further motivating its use. Similarly to F_1 scores the improvements are more prominent when evaluating only on sentences containing missing or rare senses. CMBT achieves best scores on the full test sets (all) of the *unseen* and *sample-10* setups, and a minor decrease on *big*. However, BLEU score differences are minor and they do not correlate well with F_1 improvements. As we show next, the minor differences in BLEU scores here are due to the fact that our approach mainly affects the translation of multi-sense words while leaving the translation of other words intact. It is worth pointing out that BLEU scores may not be the ideal metric in this study as they are less sensitive to word-level translation improvement as compared with F1 scores. This is similar to the findings of Arthaud et al. (2021), who showed that improving the translation of a few selected words could lead even to a slight drop in BLEU.

Analysis We manually looked at the translations of a few multi-sense words to have a better understanding of our system. We present a few examples of the typical improvements and errors we found in Table 5. In example 1 both the BWEs based system and CMBT correctly translated *bank* to *Ufer* (river bank) which shows the positive effects of the additional data. In contrast, in 2 and 3 which are examples produced under the *sample-10* and *unseen* conditions respectively, only CMBT managed to pick the right senses due to the better exploitation of the given context. All systems are incorrect in 4, however the output of CMBT (brake pedal) is related to vehicle/gas pedals (the correct sense), while the base system's output, *Beschleuniger* is more related to physics and chemical reactions, such as particle accelerator or a catalyst. We reviewed the top-5 translations given by BWEs and XLM-R for *accelerator* when it has the *gas pedal* sense in a test sentence, and found that the translations reflect the outputs of the MT systems. This shows the effectiveness of the MT system's update process and that improving CCWR based translation could lead to further improvements.

In example 5 CMBT is misled by the *mossy bank*, thus outputs the *river bank* instead of the *bench* sense in contrast to the baseline which correctly used the frequent *Bank* word. Example 6 is incorrectly translated by all systems with different

1.	SRC	*It is seen from afar sprawling along the **banks** like a cowherd taking a siesta by the water-side.*
	BASE	Es scheint aus der Ferne zu sein, wie ein Kabeljau an der Wasserseite eine Siesta nimmt.
	BWE	Es scheint aus der Ferne an den **Ufer** zu ziehen wie ein Fisch, der an der Wasserseite eine Siesta nimmt.
	CMBT	Es scheint aus der Ferne an den **Ufer** zu rasen wie ein Hirsch, der an der Wasserseite eine Siesta nimmt.
	REF	*Schon von weitem sieht man den Ort am **Ufer** lang hingestreckt liegen, wie einen Kuhhirten, der sich faulenzend am Bache hingeworfen hat.*
	GLOSS	**river bank**;
2.	SRC	*Working men, kneeling on the **banks**, washed their bare arms in the water.*
	BASE	Arbeiter, die an den **Banken** knieten, wuschen ihre bloßen Waffen im Wasser.
	BWE	Arbeitende Männer knieten an den **Banken** nieder und wuschen ihre bloßen Arme im Wasser.
	CMBT	Arbeitende Männer knieten am **Ufer** nieder und wuschen ihre bloßen Arme im Wasser.
	REF	*Arbeiter kauerten am **Ufer** und wuschen sich die Arme in der Flut.*
	GLOSS	**river bank**; **bench**;
3.	SRC	*The physician, to whom the soldiers of the **watch** had carried him at the first moment...*
	BASE	Der Arzt, zu dem ihn die Soldaten der **Uhr** im ersten Augenblick getragen hatten...
	BWE	Der Arzt, zu dem ihn die Soldaten der **Uhr** im ersten Augenblick getragen hatten...
	CMBT	Der Arzt, zu dem ihn die Soldaten der **Wache** im ersten Augenblicke getragen hatten...
	REF	*Der Heilkünstler, zu welchem die Soldaten der **Wache** ihn im ersten Augenblicke getragen...*
	GLOSS	**guard**; **timepiece**;
4.	SRC	*Try to avoid depressing the **accelerator** pedal beyond the pressure point (kickdown).*
	BASE	Versuche zu vermeiden, den **Beschleuniger**-Pedal über den Druckpunkt hinaus zu deprimieren (Kickdown).
	BWE	Versuche, den **Beschleunigerpedal** über den Druckpunkt hinaus nicht zu deprimieren (Kickdown).
	CMBT	Versuche, das **Bremspedal** über den Druckpunkt hinaus nicht zu deprimieren (Kickdown).
	REF	*Treten Sie das **Fahrpedal** möglichst nicht über den Druckpunkt durch (Kickdown).*
	GLOSS	**gas pedal**; **brake pedal**; **catalyst, (particle) accelerator**;
5.	SRC	*A lover finds his mistress asleep on a mossy **bank**;...*
	BASE	Ein Liebhaber findet seine Geliebte schlafend auf einer feuchten **Bank**;...
	BWE	Ein Liebhaber findet seine Geliebte schlafend auf einem feuchten **Bankett**;...
	CMBT	Ein Geliebter findet seine Geliebte schlafend auf einem feuchten **Ufer**;...
	REF	*Ein Liebender findet seine Geliebte auf einer moosigen **Bank** eingeschlafen;...*
	GLOSS	**bench**; **banquet**; **river bank**;
6.	SRC	*I should like to deal with one concrete point, the question of the electronic **counter**.*
	BASE	Ich möchte auf einen konkreten Punkt eingehen, die Frage des elektronischen **Gegensatzes**.
	BWE	Ich möchte mich mit einem konkreten Punkt befassen, der Frage des elektronischen **Automaten**.
	CMBT	Ich möchte auf einen konkreten Punkt eingehen, die Frage des elektronischen **Zählers**.
	REF	*Eingehen möchte ich auf einen konkreten Punkt, den Punkt der elektronischen **Schalter**.*
	GLOSS	**checkout counter**; **contrast, opposition, difference**; **vending machine**; **electricity/energy meter**;

Table 5: Example sentences highlighting the multi-sense words and their translations. For each source sentence (SRC) with given reference translation (REF) we compare the baseline (BASE) to the BWE and CMBT based systems. Word senses (GLOSS) are color coded.

errors.

Additionally, by comparing the full outputs of the systems it can be seen that our approach is non-invasive, i.e., it mostly affects the translations of multi-sense words and leaves the other parts of the sentences unchanged compared to the baseline, which is a big advantage of our approach and also explains the small BLEU differences in Table 4.

Finally, we present mined sentences based on two example source sentences containing the word bank in Table 6. The sentences indicate that our XLM-R based mining technique not only outputs the translation of the right sense but the mined sentences have similar contexts to the source sentences. This allows the fine-tuned MT system to leverage information learned from sentences that are closely related to the input sentence during translation.

6 Conclusions

In this paper we proposed CMBT, a simple and effective approach for improved rare and unseen word sense translation. It serves as a general framework that effectively exploits the context-dependent cross-lingual correspondence from a pre-trained CCWR for an MT system. We show CMBT brings significant improvements for multi-sense word translation on the English→German MuCoW test set. The improvements are the most pronounced when we directly targeted the evaluation of the difficult rare and unseen senses. As the only requirement of CMBT, on top of the parallel data necessary for the training of the MT system, is a monolingual corpus and an off-the-shelf pre-trained multilingual model, CMBT can be applied

1.	SRC	*For example, a wife learns that her husband put money in the **bank** in his name rather than in a joint account.*
	top-1	Der Abfluss bei einer Überweisung erfolgt im Zeitpunkt der Abgabe des Überweisungsauftrags an die **Bank**...
	top-1 BT	The flow of a transfer is made when the contract is delivered to the **bank**...
	top-2	Osama und Yeslam bin Laden hatten von 1990 bis 1997 ein gemeinsames Konto bei der Schweizer **Bank** UBS.
	top-2 BT	Osama and Yeslam bin Laden shared an account at the Swiss **Bank** UBS between 1990 and 1997.
2.	SRC	*At this decisive moment in Dutch history my father was positioned on the **bank** of the river Waal near the city of Nijmegen.*
	top-1	Der Highway führt nördlich am Stadtzentrum vorbei und gelangt von dort an das **Ufer** des Ontariosees.
	top-1 BT	The Highway passes north of the center of the city and then reaches the **bank** of Lake Ontario.
	top-2	Die Großstadt Pakokku liegt auf der nördlichen **Uferseite**$^{[bank-side]}$ des Irrawaddy 30 Kilometer nordöstlich von Bagan...
	top-2 BT	The big city of Pakokku is situated on the northern **bank** of Irrawaddy, 30 kilometres northeast of Bagan...

Table 6: Mining examples with XLM-R for two source sentences (SRC) containing the two senses (financial and river) of the word bank. We show the 2 highest scoring candidates and their back-translations (BT).

easily to other languages and MT systems in the future.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback and the Cambridge LMU Strategic Partnership for seed funding for this project.[9] We acknowledge Peterhouse College at University of Cambridge for funding Qianchu Liu's PhD research. The work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (No. 640550) and by the German Research Foundation (DFG; grant FR 2829/4-1) awarded to Alexander Fraser as well as by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (No. 648909) and the ERC PoC Grant MultiConvAI (No. 957356) awarded to Anna Korhonen.

References

Eneko Agirre, Oier Lopez de Lacalle, Christiane Fellbaum, Shu-K ai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. 2010. SemEval-2010 Task 17: All-Words Word Sense Disambiguation on a Specific Domain. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 75–80.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 789–798.

Farid Arthaud, Rachel Bawden, and Alexandra Birch. 2021. Few-shot learning through contextual data augmentation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1049–1062.

[9] https://www.cambridge.uni-muenchen.de

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Steven Cao, Nikita Kitaev, and Dan Klein. 2019. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2017. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. 2020. Improving cross-lingual word embeddings by meeting in the middle. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 294–304.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Associat ion for Computational Linguistics: Human Language Technologies*, pages 644–648.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5.

Denis Emelin, Ivan Titov, and Rico Sennrich. 2020. Detecting Word Sense Disambiguation Biases in Machine Translation for Model-Agnostic Adversarial Attacks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7635–7653.

Kawin Ethayarajh. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 55–65.

Matthias Huck, Viktor Hangya, and Alexander Fraser. 2019. Better OOV Translation with Bilingual Terminology Mining. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5809–5815.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681.

Frederick Liu, Han Lu, and Graham Neubig. 2018. Handling Homographs in Neural Machine Translation. In *Proceeding of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, pages 1336–1345.

Qianchu Liu, Diana Mccarthy, Ivan Vulić, and Anna Korhonen. 2019. Investigating Cross-lingual Alignment Methods for Contextualized Embeddings with Token-Level Evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, pages 33–43.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluatio n of Systems for the Semantic Analysis of Text*, pages 25–28.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linki ng. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Maria Pelevina, Nikolay Arefyev, Chris Biemann, and Alexander Panchenko. 2016. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 174–183.

Mohammad Taher Pilehvar, Jose Camacho-Collados, Roberto Navigli, and Nigel Collier. 2017. Towards a seamless integration of word senses into downstream nlp applications. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1857–1869.

Matt Post. 2019. A Call for Clarity in Reporting BLEU Scores. In *Proceedings ofthe Third Conference on Machine Translation*, pages 186–191.

Xiao Pu, Nikolaos Pappas, James Henderson, and Andrei Popescu-Belis. 2018. Integrating Weakly Supervised Word Sense Disambiguation into Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 6:635–650.

Xiao Pu, Nikolaos Pappas, and Andrei Popescu-Belis. 2017. Sense-Aware Statistical Machine Translation using Adaptive Context-Dependent Clustering. In *Proceedings of the Second Conference on Machine Translation*, pages 1–10.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW Test Suite at WMT 2019: Automatically Harvested Multilingual Contrastive Word Sense Disambiguation Test Sets for Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation*, pages 470–480.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2020. An Evaluation Benchmark for Testing the Word Sense Disambiguation Capabilities of Machine Translation Systems. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 3668–3675.

Annette Rios, Laura Mascarell, and Rico Sennrich. 2017. Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings. In *Proceedings of the Conference on Machine Translation*, pages 11–19.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96.

Kazuya Shimura, Jiyi Li, and Fumiyo Fukumoto. 2019. Text categorization by learning predominant sense of words as auxiliary task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1109–1119.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2018. An Analysis of Attention Mechanisms: The Case of Word Sense Disambiguation in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation*, pages 26–35.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2020. Encoders Help You Disambiguate Word Senses in Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1429–1435.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232.

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*, pages 5998–6008.

Ivan Vulić and Anna Korhonen. 2016. On the Role of Seed Lexicons in Learning Bilingual Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 247–257.

Yingce Xia, Xu Tan, Fei Tian, Fei Gao, Weicong Chen, Yang Fan, Linyuan Gong, Yichong Leng, Renqian Luo, Yiren Wang, et al. 2019. Microsoft Research Asia's Systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation*, pages 424–433.

Zhi Zhong and Hwee Tou Ng. 2012. Word sense disambiguation improves information retrieval. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 273–282.

A mBERT vs. XLM-R

We report the results of our initial word translation accuracy experiments using off-the-shelf mBERT and XLM-R (large) on all the multi-sense words provided by the gold MuCoW test set in Table 7. For efficiency, we randomly sampled $100K$ target sentences from Wikipedia as the candidate pool instead of the $2M$ described in Section 4.3. We take the average of the top-half layers of mBERT (top 6 layers) and XLM-R (top 12 layers) respectively when calculating word representations. We show that XLM-R performs significantly better than mBERT.

Model	$acc@1$	$acc@5$	$acc@10$
mBERT	21.40	32.16	37.29
XLM-R	**27.13**	**38.76**	**43.81**

Table 7: Comparing the translation accuracy of off-the-shelf mBERT and XLM-R on the MuCoW test set.