Towards Handling Compositionality in Low-Resource Bilingual Word Induction

Viktor Hangya hangyav@cis.lmu.de
Alexander Fraser fraser@cis.lmu.de
Center for Information and Language Processing, LMU Munich, Munich, Germany

Abstract

Bilingual word embeddings (BWEs) facilitate the translation of single source language words to single target language words. However, often a single source word must be translated using two target words. Previous approaches depend on observing the two target language words as a (frequent) bigram in a corpus. But for many languages only a small amount of written text is available, so that such "atomic" embeddings can only be built for a small number of frequent bigrams. In this paper, we extend atomic embedding based approaches to improve the 1-to-2 word translation of rare words by decomposing the representation of a source word to representations of two target words, allowing to model translations for which the required bigram was not observed in our monolingual corpora. We create a gold standard lexicon for 1-to-2 translation containing source German compounds along with their translations to two English words, and show that our approach improves performance. We also show the importance of bigrams for the downstream task of unsupervised machine translation and show small but significant BLEU score improvements with our approach. Our approach is an important first step in the direction of handling composition in BWEs, beyond simple memorization of seen bigrams.

1 Introduction

Bilingual word embeddings (BWEs) are key components in cross-lingual NLP tasks alleviating data scarcity for many languages. They can be built using source and target language monolingual corpora with either a cheap bilingual signal (Mikolov et al., 2013b; Xing et al., 2015) or no bilingual signal at all (Conneau et al., 2018; Artetxe et al., 2018a). They are applied to many downstream tasks, such as bilingual lexicon induction (BLI) (Vulić and Korhonen, 2016) and cross-lingual transfer learning (Schuster et al., 2019). Unsupervised machine translation (UMT) strongly depends on BWEs. Meaningful translations can be generated without any bilingual signal by using BWEs to translate words 1-to-1.

However, many words in a given language are the composition of multiple smaller lexical units which are expressed individually in other languages, such as the German compound *Waschmaschine* $\rightarrow$ *washing machine*. Using the idea of atomic embeddings for frequent bigrams (Mikolov et al., 2013c) previous work proposed 1-to-2 word translations achieving significant improvements in UMT (Lample et al., 2018b; Artetxe et al., 2018b). These approaches start by learning individual vector representations for frequent bigrams using their monolingual contexts, effectively treating them as if they were single words. After the projection of the learned monolingual spaces to BWEs, these special bigrams can also be translated or serve as translation candidates in exactly the same way as unigrams. On the other hand, these approaches need to learn these special atomic embeddings in advance, requiring them to be frequently observed to be learned well. This assumes the availability of large monolingual corpora which are

not available for low-resource languages. Relying only on atomic embeddings fails if the target bigrams are out-of-vocabulary or infrequent. For example, if the English bigram in *Cocabauern* $\rightarrow$ *coca farmers* was not seen in the target corpus, then its embedding will be unknown. Thus it cannot serve as a translation candidate. Furthermore, if it was seen but it is infrequent, its representation will be poor. We therefore have a high chance of translating *Cocabauern* to a similar but more frequent English bigram (e.g. *corn farmers*). This problem was partially addressed in (Del et al., 2018) by generating bigram representations using a linear composition based on the embeddings of the two words in a bigram. This way they are able to generate better embeddings for infrequent bigrams by relying on their frequent unigram components. On the other hand, their approach still suffers from the OOV issue because they are only able to compute a limited number of compositions, which they restrict to observed bigrams.

To overcome these problems we extend atomic embeddings in order to improve translation of rare words by decomposing source language unigrams to two target words. Given the embedding of a unigram (*Cocabauern*) we infer two word vectors which we decode to a bigram (*coca farmers*) by looking for the nearest neighbors of the generated vectors. This way we omit the need for atomic bigram representations on the target side learned in advance and allow the generation of previously unseen bigrams. We employ a multi-layer neural network relying only on unsupervised BWEs and generate cross-lingual training examples using atomic embeddings and back-translation (Sennrich et al., 2016). In addition, we create monolingual examples having a target language bigram as the input and its unigram components as the output, i.e., these examples teach the system to map the atomic embedding of a bigram like "corn_farmers" to the two representations of "corn" and "farmers", which monolingually mimics the cross-lingual translation task. Similarly to auto-encoding based approach, we can generate these monolingual examples easily which are useful for the bilingual task as well since the BWEs on which our system relies represent source and target language words with similar vectors.

We test our system on the new task of bilingual phrase lexicon induction (BPLI), i.e., translating single words to phrases, which we propose here. We use German compound words and their English translations as the test lexicon, since they behave well in 1-to-2 translation, serving as a good starting point for our experiments. We focus on 1-to-2 translation only since they have the highest impact on UMT quality (Lample et al., 2018b). We simulate De-En as low-resource by using a large amount of monolingual data only on the source side (De) while testing various data sizes for the target (En). We show that by combining our proposed approach with atomic embeddings the performance on the BPLI task can be improved. Our analysis shows that mapping a source word to atomically embedded bigrams has high performance when translating frequent source words, while the decomposition of a source word to two target word representations works well when translating infrequent source words. In addition, we show the importance of the system for UMT by including our approach in the pipeline of the UMT system of Artetxe et al. (2018b) and show its positive effects on translation quality.

2 Related Work

Bilingual word embeddings became popular resources for many cross-lingual NLP tasks since they allow the transfer of knowledge from a source language to a target language. Various approaches were proposed. Gouws et al. (2015) rely on parallel corpora, while others create artificial cross-lingual corpora using seed lexicons or document alignments (Vulić and Moens, 2015; Duong et al., 2016). Following Mikolov et al. (2013b), many authors map monolingual word embeddings (MWEs) into shared bilingual spaces (Faruqui and Dyer, 2014; Xing et al., 2015) because only a weak bilingual signal in the form of a seed lexicon is needed. Bilingual Lexicon Induction (BLI) is often used as the intrinsic evaluation of BWE spaces (Mikolov et al., 2013b; Vulić and Korhonen, 2016), where the task is to translate individual source language

words to a single target word (1-to-1). Most approaches rely on cosine similarity of word embeddings by predicting the top-1 or top-5 most similar words as translations. In order to evaluate our approach intrinsically, we created a test lexicon as is usually done in BLI, but our test lexicon has single words and their bigram translations.

Recent work showed that building BWEs is possible without any bilingual signal. Adversarial training was used to rotate the source space to match the target in order to extract an initial seed lexicon which is used to fine-tune the projection (Conneau et al., 2018). Others used word neighborhood information to create the initial mapping (Artetxe et al., 2018a; Alvarez-Melis and Jaakkola, 2018). All these works led to the possibility of building MT systems without parallel data which are based on the word translation capabilities of unsupervised BWEs and back-translation of large monolingual data (Sennrich et al., 2016). Systems based on both neural network approaches (Lample et al., 2018a; Artetxe et al., 2018c; Yang et al., 2018) and phrase-based SMT (Lample et al., 2018b; Artetxe et al., 2018b) were proposed. We evaluate our approach on the downstream task of UMT as well by extending the approach of (Artetxe et al., 2018b), showing translation quality improvements.

An important step of statistical UMT systems which gives significant performance improvement is learning atomic representations for bigrams. Statistical UMT allows 1-to-2, 2-to-1 and even 2-to-2 translations. Mikolov et al. (2013c) showed that good quality embeddings can be learned by mining frequent n-grams in monolingual corpora using co-occurrence statistics and learning their representation in the same way as other vocabulary entries. The approach was improved in (Artetxe et al., 2018b) by keeping unigram invariance while learning n-gram embeddings. We give more details about this approach in the following section since we rely on this system in our experiments. On the other hand, the problem of these approaches is that a large amount of monolingual data is required in order to mine frequent bigrams and to learn good quality embeddings for them. Del et al. (2018) alleviated the problem by inferring bigram embeddings by composing the representations of their unigram components. Various composition functions were tested, such as simple addition of vectors or learning a linear projection. Following the work of Yazdani et al. (2015), they use atomic bigram embeddings and the representations of their unigram components as training samples to learn the composition function. However it is not feasible to compose each pair of unigrams in the vocabulary due to their large number, thus the approach considers only those bigram candidates which occur in the input monolingual corpora, typically leading to OOV target bigrams in the case of rare source words. Our approach alleviates these problems by extending previous systems with a decomposition based module for better rare word translation which is able to generate previously unseen bigrams as well.

3 Approach

As mentioned above in the approach of Del et al. (2018) the embeddings of target language bigrams are generated using a composition mechanism which are used as target candidate translations for source words. Since they show that the composition of target language unigrams can be used for 1-to-2 translation, we can assume that source word embeddings encode the meaning of its components. Based on this intuition, we follow a reverse approach where we decompose source word representations into two vectors which we decode into bigrams using the target language vocabulary. The advantage of this approach is that it does not require a predefined list of bigrams. Kumar and Tsvetkov (2019) proposed a supervised MT system which generates word embeddings, instead of word indices, on the output which are then decoded into sentences using beam-search and a word embedding model. Similarly, we generate vectors as an intermediate step instead of word indices directly since our training lexicon covers only a subset of the target language vocabulary (a logit layer would be able to predict only the words in this subset).

We use a two layer feed-forward neural network[1] as the decoder which takes the embedding of the source word as input:

$$[y_{t_1}, y_{t_2}] = W_2 * ReLU(W_1 * drop(x_s)) \tag{1}$$

where x_s is the input while $[y_{t_1}, y_{t_2}]$ is the concatenated output word vectors, W_i are network parameters and $ReLU, drop$ are the non-linearity and dropout functions respectively. As training objective we minimize the mean cosine distance between predicted and gold translations in the training lexicon:

$$\underset{W_1,W_2}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} d([y^i_{t_1}, y^i_{t_2}], [\hat{y}^i_{t_1}, \hat{y}^i_{t_2}]) \tag{2}$$

where $\hat{y}^i_{t_i}$ is the embedding of gold translations and $d(\cdot, \cdot)$ is the cosine distance of the vectors. The probability of a bigram $[w_{t_1}, w_{t_2}]$ is given by:

$$P([w_{t_1}, w_{t_2}] \mid [y_{t_1}, y_{t_2}]) = \frac{s(y_{t_1}, x_{w_{t_1}}) * s(y_{t_2}, x_{w_{t_2}})}{\sum_{w_1,w_2 \in V_t} s(y_{t_1}, x_{w_1}) * s(y_{t_2}, x_{w_2})} \tag{3}$$

where x_{w_i} is the embedding of word w_i, V_t is the target language unigram vocabulary and $s(\cdot, \cdot)$ is the cosine similarity of vectors. Since w_{t_i} can be any unigram in the target vocabulary we only consider the 100 most similar words compared to y_{t_1} and y_{t_2} as possible values for w_{t_1} and w_{t_2} respectively, based on their cosine similarity. By assuming that the rest of the words are irrelevant for the decoding of the output we significantly reduce the search space. In addition, we omit the normalization term during decoding for efficiency.

Atomic Embeddings As the basis of our system we use atomic embeddings of bigrams. We learn BWEs containing such embeddings using the method of (Artetxe et al., 2018b). In the first step the system learns source and target language MWEs containing frequent n-grams based on co-occurrence statistics (Mikolov et al., 2013c). Opposed to previous approaches, which treated n-grams the same way as unigrams, Artetxe et al. (2018b) modified word2vec skip-gram (Mikolov et al., 2013a) in a way that it learns n-gram embeddings keeping unigram invariance, i.e., resulting unigram embeddings are the same as when there are no longer phrases in the vocabulary. This is achieved by only updating n-gram embeddings but not context embeddings when a training example has an n-gram as the center word. The projection of MWEs to a BWE space is done using the unsupervised *VecMap* system (Artetxe et al., 2018a). The method builds an initial mapping relying on intra-lingual similarity distribution of embeddings and iteratively improves the projection through self-learning without any bilingual signal.

Training lexicons Since our aim is to apply our system to setups where no bilingual signal is available, we generate training lexicons automatically, containing source→target translation pairs, relying only on unsupervised BWEs. We build multiple lexicon variations containing either cross-lingual or monolingual training examples. As the quality of atomic embeddings are good in case of frequent bigrams, we use them to learn basic decomposition and to generalize decomposition to rare and OOV bigrams as well. When training our system the source side of each example is represented by a single, while the target by two individual vectors.

[1] Simple linear projection resulted in lower performance.

s2t We build a cross-lingual lexicon containing source language unigrams having target language bigram translations, e.g., *Waschmaschine* → *washing machine*. For simplicity, we translate target language bigrams back to the source language using BLI, i.e., by taking to most similar unigram or bigram as the translation based on their vector similarity. We filter pairs which have a bigram on the source side, since we are only interested in unigram→bigram pairs. Finally, we retain only those pairs which have at least 0.2 similarity in order to have a good quality lexicon.

t2t We create a monolingual lexicon as well, containing the same target language bigrams on both source and target sides. We use the atomic bigram embeddings to decompose them to the individual word vectors of their unigram components, e.g., *washing_machine* → *washing machine*, similarity to (Yazdani et al., 2015; Del et al., 2018). Since we rely on BWEs, i.e., words of the two languages are represented in a shared space, this lexicon can be used in the same way as *s2t* by considering the target language bigram vectors on the source side as the noisy representation of the source language bigrams. We take the list of target language bigrams from the BWE vocabulary for this lexicon but clean it by filtering any bigram which has a component that is a stopword, a punctuation mark, a digit, shorter that 4 characters or if its POS sequence[2] is not composed of two nouns, a noun and an adjective or does not start with a verb (gerund, present or past participle) and ends with a noun. We note that the POS patterns reflect German compound composition but they can easily be extended to generalize our approach in future work. In addition, we keep only the most frequent 10K bigrams (or less if 10K is not available). We apply the same filtering to the following two lexicons as well except that we vary the number of the retained entries as described below.

s2t-avg We found that having too many bigrams in the BWE vocabulary when not enough monolingual data is available on the target side leads to bad quality atomic embeddings. This can be alleviated by learning embeddings for frequent bigrams only, but this leads to small *s2t* and *t2t* lexicons. To overcome this problem, we mine additional less frequent bigrams from the target language corpus without learning atomic embeddings for them due to their low frequency. We create a lexicon using these bigrams and by back-translation (Sennrich et al., 2016) applying BLI on their unigram components individually. We then take the average of the two resulting source language unigrams' representations as the source side bigram vector, since atomic embeddings are not available.

t2t-avg We generate a monolingual lexicon in a similar way. We use the target language bigrams from *s2t-avg* and take the average target language representations of the bigrams' components as the source side vector instead of back-translating them. Since learning atomic embeddings is better than averaging, we use only the most frequent 1K bigrams.

Because not all lexicons are equally useful for the training of our system due to their quality, we use them sequentially by performing 10 epochs each on *t2t-avg*, *s2t-avg*, and *t2t* respectively in this order, and then we run 70 epochs on *s2t*. We note that to improve the precision of these lexicons we used CSLS (Conneau et al., 2018) instead of cosine similarity for BLI.

Ensembling In order to improve 1-to-2 translation of rare words we extend atomic embeddings with the decomposition approach by ensembling the outputs of the two modules. We define $T^B_{w_s}$ a set of 100 target language bigrams that are most similar to the input word w_s based on CSLS similarity of embeddings in a given BWE space. Similarly, $T^D_{w_s}$ is a set of most probable 100 translations predicted by our approach. We calculate the ensemble score for each w_s and $t_{w_s} \in T^B_{w_s} \cup T^D_{w_s}$ as:

[2]We use *spaCy* for POS tagging `https://spacy.io`

$$S_E(w_s, t_{w_s}) = \sum_{m \in B,D} \lambda_m * S_m(w_s, t_{w_s}), \qquad S_m = \frac{s_m(w_s, t_{w_s})}{\sum_{t \in T^B_{w_s} \cup T^D_{w_s}} s_m(w_s, t)} \tag{4}$$

where $s_B(\cdot,\cdot)$ is the CSLS similarity between its arguments' representations, $s_D(\cdot,\cdot)$ is the prediction score of the decomposition module and λ_B, λ_D are the weights of the two methods respectively. If a given (w_s, t) is not in $T^m_{w_s}$ we set $s_m(w_s, t)$ to 10^{-6}.

Parameters The parameters in our experiments are the following: we use 300 dimensional word embeddings, hidden layer of 1000 and dropout probability 0.2 in our decoder, batch size of 32 over 100 training epochs without early stopping. The learning rate of the *Adam* optimizer is 0.001 initially which we multiply by 0.1 in every 10^{th} epoch when using training examples of *s2t* lexicon. We used the development set of the created BPLI lexicon (see below) for tuning the network parameters and the ensemble weights. We implemented our system in PyTorch (Paszke et al., 2019).

We note that previous work learned trigram embeddings as well for UMT systems (Lample et al., 2018b; Artetxe et al., 2018b). We only focus on bigrams since they have the most impact, while higher n-grams have only marginal improvements, as shown by Lample et al. (2018b). However, based on the lexicon generation and equation 1, the extension of our system from bigrams to longer n-grams, which we leave for future work, is technically straight-forward. One needs to generate lexicons for longer n-grams as well, while extending the number of vectors predicted by the decoder. To allow for variable n-gram length the introduction of a *PAD* token is necessary.

4 Evaluation

4.1 Bilingual Phrase Lexicon Induction

We introduce a novel test lexicon containing German compounds and their English translations for the evaluation[3]. We focus on compounds in this work since they behave well in 1-to-2 translations. We will discuss future generalization possibilities of the approach, such as non-compound inputs, at the end of the paper. We take the source compounds from the work of Fritzinger and Fraser (2010) which was created to test German compound splitting accuracy. Since the dataset does not contain English pairs of the words we automatically translated them using Google Translate. Some of the compounds were translated 1-to-1, e.g., $Ackerland \rightarrow farmland$, we filtered them out which resulted in 661 pairs. Besides bigrams on the target side, the lexicon contains 3-grams and 4-grams as well, making up around 10% and 1% of the examples respectively. Since our current system only translates to bigrams, longer phrases count as errors in our evaluation, but we kept these entries in the lexicon to allow future comparison. We use half of the lexicon for parameter tuning and the other half for testing.

As the baseline we learn BWEs containing atomic bigram embeddings using the method of (Artetxe et al., 2018b) described above and perform BLI. For both the baseline (*atomic*) and our system (*ensemble*), we only allow bigram outputs. To build BWEs we use the same monolingual data used in (Artetxe et al., 2018b) which contains German ($89.6M$) and English ($90.2M$) news crawl sentences between 2007 and 2013 released by WMT14. We simulate low-resource settings by decreasing the amount of available sentences on the target side ($all, 1M, 500K$ and $250K$) but keeping the full dataset on the source. Our experiments showed that having too many atomic n-gram embeddings when we have only a low number of sentences results in low quality BWEs which can be improved by decreasing the number of n-gram embeddings. Based

[3]The dataset is available at: `https://www.cis.lmu.de/~hangyav/data/BPLI.zip`

on this and the setup of (Artetxe et al., 2018b), we build BWEs containing $200K$ most frequent embeddings for unigrams and $400K$ most frequent bigrams and trigrams on both source and target side, except in the case of $250K$ sentences where we used $10K$. We note that we also tried to experiment with only $100K$ sentences in the monolingual data but even with only unigrams in the vocabulary we couldn't build functional BWEs.

4.2 Unsupervised Machine Translation

We evaluate our system extrinsically as well on the task of unsupervised MT on the WMT news translation shared task test data from 2014 and 2016, similarly to previous work on unsupervised MT (Lample et al., 2018b; Artetxe et al., 2018b). We additionally evaluate on WMT 2019 (Barrault et al., 2019).

To examine the effects of BWEs on MT, we extend the statistical UMT system of (Artetxe et al., 2018b) which strongly relies on BWEs. By default it is comprised of 5 consecutive steps. In step 1 it builds MWEs containing atomic representations for 1-, 2- and 3-grams and projects them to shared BWE spaces in step 2 with the same method as the basis of our system for the BPLI experiments described above. The Moses statistical MT system is used as the translation system (Koehn et al., 2007) which requires two components in this case: a language model and a phrase-table. The former is built with KenLM (Heafield, 2011) while the latter contains phrase translation pairs and their scores for each source word and the 100 most similar target words which are calculated using cosine similarity in step 3. The weights of the two components are tuned in step 4 on a synthetic parallel corpus generated through back-translation and MERT is applied as the tuning algorithm. Finally, in step 5 the system is iteratively refined (for 3 iterations) using back-translation of monolingual data. We use this off-the-shelf system as one of our baselines which we call *atomic* since it uses atomic n-gram embeddings. In addition, we run the same system but without atomic representations for 2- and 3-grams, i.e., with only unigram word embeddings (*unigram*).

Extended UMT To plug our system into the pipeline we extend the generated phrase-table in step 3 of the atomic baseline to show the additional effects of the bigrams generated by our approach. We filter words from the source language vocabulary which are longer than the average character length in our BPLI test lexicon (13) with the aim to gather compound words automatically. We predict top-100 translations of these words and create a phrase-table using the prediction score of the decomposition module as the translation probability. We then merge it with the baseline phrase-table by taking their union. More precisely, each entry has 5 scores, 4 coming from the baseline and 1 from our extension. In addition, if a source-target pair is missing in one of the tables we set the related values to 10^{-6}. We only extend the De→En phrase-table but as discussed below it also affects the En→De direction as well due to back-translation. We create a secondary phrase-table with the decomposition module instead of generating just one with the extended atomic embeddings because this way the tuning procedure in step 4 is able to tune parameters for the two phrase-tables more precisely. Other steps in the original pipeline are unchanged. We refer to this system as *extended*.

5 Results

5.1 Bilingual Phrase Lexicon Induction

To evaluate BPLI we report top-5 accuracy (acc_5) on lowercased words, as is done in most work on BLI (Mikolov et al., 2013b). Some of the source words in the test lexicon are OOVs, i.e., their embeddings are unknown. We filtered them since no system is able to predict their translations. We show results in Table 1 comparing the baseline *atomic* embeddings and its extension with the decomposition module (*ensemble*). As mentioned earlier we use various target corpora sizes.

	atomic	ensemble	λ_B
250K	12.12	15.06	0.50
500K	34.34	37.35	0.46
1M	43.98	46.99	0.58
all	65.06	65.06	0.70

Table 1: Acc_5 results on the BPLI task in percentage points of the baseline (atomic) and our proposed approach (ensemble). The weight of BWE based similarities are given by λ_B ($\lambda_D = 1.0 - \lambda_B$).

	atomic			decomposition		
	all	≤ 100	≤ 10	all	≤ 100	≤ 10
250K	12.12	2.04	0.00	13.86	8.11	1.15
500K	34.34	27.54	1.52	19.88	15.22	4.55
1M	43.98	31.62	2.13	21.69	14.53	4.26
all	65.06	0.00	0.00	25.90	0.00	0.00

Table 2: Comparison of the two modules on various target language data sizes (1^{st} column) and limited test lexicons containing *all* source words or only those which have a gold translation with frequencies ≤ 100 and ≤ 10 respectively (2^{nd} row).

	all	≤ 100	≤ 10
250K	325	306	235
500K	325	295	200
1M	325	269	162
all	325	69	33

Table 3: Number of test lexicon entries along varying target language corpora sizes (1^{st} column) and bigram frequencies (1^{nd} row).

When only a small amount of target language data is available the quality of atomic embeddings decreases due to a larger number of rare words. By combining the decomposition module with the atomic embeddings the performance increased, especially in case of the low data size setups. As the target language corpus grows there are less rare words that have to be translated 1-to-2, thus the difference between the two systems decreases but even at 1M sentences the difference is significant. No improvements were achieved for *all* in terms of acc_5 since the data size is large ($90.2M$). On the other hand, by looking at acc_1 our system performed 0.6% better. We found ensembling weights to be best when the two models contribute about equally as shown in Table 1.

Atomic vs. Decomposition In Table 2 we compare the two modules of our approach to depict their performance differences. Other than the limited target corpus sizes we create setups where the test BPLI lexicon contains only those compounds which have a gold translation with frequency at most 100 and 10 respectively. Note that the number of entries in these lexicons is changing with the target corpus size which we show in table 3. This is because as the training data size increases the identity of the limited frequency translations changes.

The results show that the decomposition module by itself performs better than atomic embeddings if there is not much data. It works even better in contrast with the atomic system when looking at low frequency bigrams only which clearly shows the advantage of our approach in

	250K	500K	1M	all
s2t	3.61	13.25	15.66	26.51
s2t+t2t	4.82	18.67	15.66	24.70
all 4	13.86	19.88	21.69	25.90
s2t-avg+t2t-avg	9.64	9.04	13.25	14.46

Table 4: Ablation study on the effect of the 4 generated training lexicons on the decomposition module. The 1^{st} column shows the used lexicons. Results are on the full test lexicon with various target language corpora sizes.

	250K	500K	1M	all
s2t	237	949	2,210	14,656
t2t	267	10,000	10,000	10,000
s2t-avg	11,963	23,249	40,818	36,988
t2t-avg	1,000	1,000	1,000	1,000

Table 5: Generated training lexicon sizes in various target language corpora sizes.

low-resource cases. For both $500K$ and $1M$ atomic performs better on the full test lexicon which shows the good quality of atomic embeddings when the data size is large. On the other hand, on frequency ≤ 10 our approach performs better even for $1M$ showing that it can generalize to low-frequency test cases better which explains the good performance of the joint model in Table 1. Furthermore, both systems achieve 0% accuracy on the limited frequency test sets taken from the *all* scenario because these test sets have only a few but very low frequency words which have low quality embeddings, thus making their translation difficult.

Ablation study We performed an ablation study to analyze the contribution of the generated training lexicons on the decomposition module. We show results using only atomic embeddings on the source side (s2t and s2t+t2t) and no atomic embeddings (s2t-avg+t2t-avg) in Table 4. The former aims at analyzing the effect of having a small number of good quality pairs (training lexicon sizes are shown in Table 5) while the latter gives an intuition of achievable performance when atomic embeddings are unavailable. Results show that the s2t set is the most important, since as its size grows the additional improvements coming from the other sets are decreasing. Note, that in case of using all target language data the best results were achieved when using only s2t. On the other hand, in the lower resource setups the other sets are essential in achieving good performance, especially in case of $250K$ when both s2t and t2t are small due to a small number of frequent bigrams.

5.2 Unsupervised Machine Translation

We compare the performance of using standard BWEs (*unigram*), with the *atomic* approach and our extension of the atomic approach (*extended*) in terms of BLEU in Table 6. We show results of step 4 (parameter tuning) and step 5 (iterative refinement). Our results show that relying on unigram BWEs (as was done for NMT in previous work) performs poorly. Comparing the unigram with the atomic and extended variants it can be seen that using bigram embeddings to perform initial word translation leads to significant improvements in both step 4 and step 5, which is caused by the translation possibility of source words to multiple target words. In addition, we achieved further improvements with the extension of the atomic system with the decomposition module by having better translation entries in the phrase table for rare words. Just as in the BPLI experiments, the combination of atomic and decomposed bigrams has posi-

			wmt14		wmt16		wmt19	
			De-En	En-De	De-En	En-De	De-En	En-De
250K	step4	unigram	3.50	3.30	5.00	4.70	3.50	3.60
		atomic	5.50	4.80	6.70	6.60	4.80	5.80
		extended	5.60	4.90	6.80	6.60	4.90	5.90
	step5	unigram	8.70	7.00	11.60	9.70	10.00	9.20
		atomic	9.90	8.20	13.00	10.90	10.60	10.70
		extended	10.10	8.30	13.40	11.30	11.10	10.80
500K	step4	unigram	5.60	3.80	7.80	5.40	5.80	4.70
		atomic	7.10	5.80	8.50	7.40	6.30	6.30
		extended	7.40	5.90	9.00	7.60	6.80	6.40
	step5	unigram	9.90	8.40	13.50	11.90	11.30	11.20
		atomic	11.80	9.60	15.30	13.10	11.90	12.20
		extended	12.00	9.50	15.40	13.10	12.60	11.70
1M	step4	unigram	7.90	5.30	10.20	7.30	7.60	6.60
		atomic	10.20	7.80	12.70	10.00	9.80	8.60
		extended	10.20	8.00	12.80	10.30	9.90	8.80
	step5	unigram	11.00	9.20	14.40	12.60	11.90	12.00
		atomic	12.70	10.70	16.50	14.10	12.90	12.80
		extended	12.80	10.30	16.50	14.10	13.00	12.60
all	step4	unigram	9.20	5.50	11.70	7.40	8.40	6.40
		atomic	14.93	10.71	18.34	13.46	13.01	11.90
		extended	15.00	11.00	18.60	13.70	13.20	12.40
	step5	unigram	12.70	10.60	17.40	14.40	12.20	13.50
		atomic	17.04	13.39	22.28	17.43	15.70	14.80
		extended	17.40	13.70	22.40	17.70	16.60	15.30

Table 6: BLEU scores comparing the two baseline (with and without n-gram embeddings) and the extended UMT systems. The first column shows the number of sentences in the target language monolingual data used to build BWEs. Step 4 and 5 are the parameter tuning and iterative refinement steps in the UMT training pipeline.

tive effects not only in the low-resource but higher resource cases as well and also when using the full target language dataset. This shows that our system generates useful bigrams, other than those in the BWE vocabulary, which are getting picked by the language model. Improvements can be seen in case of both steps 4 and 5 meaning that the parameter tuning can decide how to trade off the weighting of the atomic embeddings based phrase-table and the decomposition based phrase-tables and there are positive effects during the refinement steps as well. In addition, improvements can be seen in case of the En-De translation direction even though we only extend the De-En phrase table. Since the initial De-En model in step 4 is improved by the extension, it affects the opposite direction as well due to the use of the initial De-En model during back-translation.

6 Conclusion

Unigram BWEs do not model important 1-to-2 word translations. We show the gain for using atomic embeddings for bigrams in order to perform 1-to-2 word translation but this approach can only be applied when there is a large amount of monolingual data available. Our new decomposition based approach for modeling 1-to-2 translation of rare words directly translates source unigram embeddings to target language bigrams allowing us to predict both rare and

OOV bigrams. We proposed a system combining the advantages of atomic embeddings and decomposition which we tested intrinsically on 1-to-2 BLI of German compounds for which we created a new test set. We release our BPLI dataset for further development. We showed improved performance compared to just using atomic embeddings even in less resource-poor setups. Furthermore, our analysis showed that on rare words the decomposition based translation alone outperforms atomic embeddings, further motivating their joint use. We also showed that bigrams are important for statistical UMT systems and that by plugging our system into an UMT pipeline we achieved better performance compared to an off-the-shelf system on this extrinsic task as well.

In this paper we presented first steps towards handling compositionality in BLI and UMT. To be able to extend UMT systems by our approach in general, not only for compound translations as we have done in our preliminary UMT experiment here, the length of the output translations has to be decided dynamically. The proposed architecture is compatible with such a setup with the introduction of a padding token and training lexicon containing unigrams as well but further experimentation is required. The translation of non-compound words, such as when target language functional words are expressed with pre-, in- or suffixes in the source (e.g. $kedd + \underline{en} \rightarrow on\ Tuesday$ in Hungarian or $ver + \underline{me} + mek \rightarrow not\ to\_give$ in Turkish), should be investigated in future work as well.

Acknowledgments

We would like to thank Leonie Weißweiler for her contribution to the construction of the BPLI lexicon and the anonymous reviewers for their valuable input. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 640550).

References

Alvarez-Melis, D. and Jaakkola, T. (2018). Gromov-Wasserstein Alignment of Word Embedding Spaces. In *Proc. EMNLP*, pages 1881–1890.

Artetxe, M., Labaka, G., and Agirre, E. (2018a). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proc. ACL*, pages 789–798.

Artetxe, M., Labaka, G., and Agirre, E. (2018b). Unsupervised Statistical Machine Translation. In *Proc. EMNLP*, pages 3632–3642.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018c). Unsupervised Neural Machine Translation. In *Proc. ICLR*.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 Conference on Machine Translation (WMT19). In *Proc. WMT*, pages 1–61.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2018). Word Translation Without Parallel Data. In *Proc. ICLR*, pages 1–14.

Del, M., Tättar, A., and Fishel, M. (2018). Phrase-based Unsupervised Machine Translation with Compositional Phrase Embeddings. In *Proc. WMT*, pages 361–367.

Duong, L., Kanayama, H., Ma, T., Bird, S., and Cohn, T. (2016). Learning crosslingual word embeddings without bilingual corpora. In *Proc. EMNLP*, pages 1285–1295.

Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In *Proc. EACL*, pages 462–471.

Fritzinger, F. and Fraser, A. (2010). How to Avoid Burning Ducks: Combining Linguistic Analysis and Corpus Statistics for German Compound Processing. In *Proc. WMT*, pages 224–234.

Gouws, S., Bengio, Y., and Corrado, G. (2015). Bilbowa: Fast bilingual distributed representations without word alignments. In *Proc. ICML*.

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In *Proc. EMNLP*, pages 187–197.

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 *Proc. ACL*, pages 177–180.

Kumar, S. and Tsvetkov, Y. (2019). Von Mises-Fisher Loss for Training Sequence to Sequence Models with Continuous Outputs. In *Proc. ICLR*.

Lample, G., Denoyer, L., and Ranzato, M. (2018a). Unsupervised Machine Translation Using Monolingual Corpora Only. In *Proc. ICLR*.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. (2018b). Phrase-Based & Neural Unsupervised Machine Translation. In *Proc. EMNLP*, pages 5039–5049.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proc. ICLR*.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting Similarities among Languages for Machine Translation. *CoRR*, abs/1309.4.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013c). Distributed Representations of Words and Phrases and their Compositionality. In *Proc. NIPS*, pages 3111–3119.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, 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 *Proc. NeurIPS*, pages 8024–8035.

Schuster, T., Ram, O., Barzilay, R., and Globerson, A. (2019). Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing. In *Proc. NAACL-HLT*, pages 1599–1613.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving Neural Machine Translation Models with Monolingual Data. In *Proc. ACL*, pages 86–96.

Vulić, I. and Korhonen, A. (2016). On the Role of Seed Lexicons in Learning Bilingual Word Embeddings. In *Proc. ACL*, pages 247–257.

Vulić, I. and Moens, M. (2015). Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proc. ACL*, pages 719–725.

Xing, C., Wang, D., Liu, C., and Lin, Y. (2015). Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. In *Proc. NAACL-HLT*, pages 1006–1011.

Yang, Z., Chen, W., Wang, F., and Xu, B. (2018). Unsupervised Neural Machine Translation with Weight Sharing. In *Proc. ACL*, pages 46–55.

Yazdani, M., Farahmand, M., and Henderson, J. (2015). Learning Semantic Composition to Detect Non-compositionality of Multiword Expressions. In *Proc. EMNLP*, pages 1733–1742.