Shared-Private Bilingual Word Embeddings for Neural Machine Translation

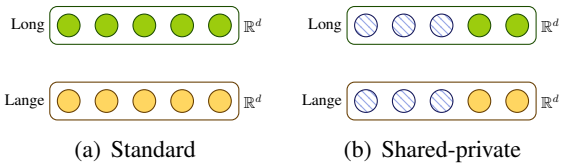
Xuebo Liu[†] **Derek F. Wong**[†*] **Yang Liu**[‡] **Lidia S. Chao**[†] **Tong Xiao**[§] **Jingbo Zhu**[§]
[†]NLP[2]CT Lab / Department of Computer and Information Science, University of Macau, Macau
[‡]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[§]Northeastern University, Shenyang, China
nlp2ct.xuebo@gmail.com, {derekfw,lidiasc}@um.edu.mo,
liuyang2011@tsinghua.edu.cn, {xiaotong,zhujingbo}@mail.neu.edu.cn

Abstract

Word embedding is central to neural machine translation (NMT), which has attracted intensive research interest in recent years. In NMT, the source embedding plays the role of the entrance while the target embedding acts as the terminal. These layers occupy most of the model parameters for representation learning. Furthermore, they indirectly interface via a soft-attention mechanism, which makes them comparatively isolated. In this paper, we propose *shared-private* bilingual word embeddings, which give a closer relationship between the source and target embeddings, and which also reduce the number of model parameters. For similar source and target words, their embeddings tend to share a part of the features and they cooperatively learn these common representation units. Experiments on 5 language pairs belonging to 6 different language families and written in 5 different alphabets demonstrate that the proposed model provides a significant performance boost over the strong baselines with dramatically fewer model parameters.

1 Introduction

With the introduction of ever more powerful architectures, neural machine translation (NMT) has become the most promising machine translation method (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015). For word representation, different architectures—including, but not limited to, recurrence-based (Chen et al., 2018), convolution-based (Gehring et al., 2017) and transformation-based (Vaswani et al., 2017) NMT models—have been taking advantage of the distributed word embeddings to capture the syntactic and semantic properties of words (Turian et al., 2010).



(a) Standard (b) Shared-private

Figure 1: Comparison between (a) standard word embeddings and (b) shared-private word embeddings. In (a), the English word "Long" and the German word "Lange", which have similar lexical meanings, are represented by two private d-dimension vectors. While in (b), the two word embeddings are made up of two parts, indicating the shared (lined nodes) and the private (unlined nodes) features. This enables the two words to make use of common representation units, leading to a closer relationship between them.

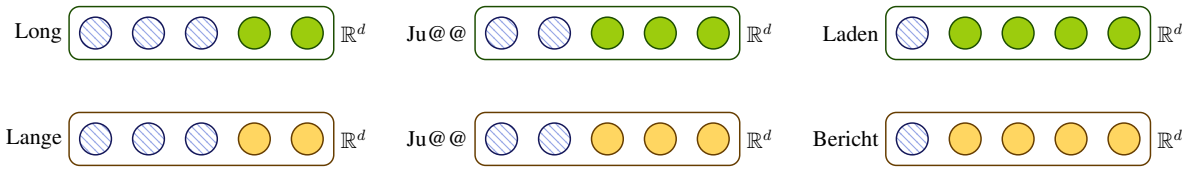
NMT usually utilizes three matrices to represent source embeddings, target input embeddings, and target output embeddings (also known as pre-softmax weight), respectively. These embeddings occupy most of the model parameters, which constrains the improvements of NMT because the recent methods become increasingly memory-hungry (Vaswani et al., 2017; Chen et al., 2018).[1] Even though converting words into subword units (Sennrich et al., 2016b), nearly 55% of model parameters are used for word representation in the Transformer model (Vaswani et al., 2017).

To overcome this difficulty, several methods are proposed to reduce the parameters used for word representation of NMT. Press and Wolf (2017) propose two weight tying (WT) methods, called decoder WT and three-way WT, to substantially reduce the parameters of the word embeddings. Decoder WT ties the target input embedding and target output embedding, which has become the new *de facto* standard of practical NMT (Sen-

*Corresponding author

[1]For the purpose of smoothing gradients, a very large batch size is needed during training.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3613–3622
Florence, Italy, July 28 - August 2, 2019.



(a) Similar lexical meaning (b) Same word form (c) Unrelated

Figure 2: Shared-private bilingual word embeddings perform between the source and target words or sub-words (a) with similar lexical meaning, (b) with same word form, and (c) without any relationship. Different sharing mechanisms are adapted into different relationship categories. This strikes the right balance between capturing monolingual and bilingual characteristics. The closeness of relationship decides the portion of features to be used for sharing. Words with similar lexical meaning tend to share more features, followed by the words with the same word form, and then the unrelated words, as illustrated by the lined nodes.

nrich et al., 2017). Three-way WT uses only one matrix to represent the three word embeddings, where the source and target words that have the same word form tend to share a word vector. This method can also be adapted to sub-word NMT with a shared source-target sub-word vocabulary and it performs well in language pairs with many of the same characters, such as English-German and English-French (Vaswani et al., 2017). Unfortunately, this method is not applicable to languages that are written in different alphabets, such as Chinese-English (Hassan et al., 2018).

Another challenge facing the source and target word embeddings of NMT is the lack of interactions. This degrades the attention performance, leading to some unaligned translations that hurt the translation quality. Hence, Kuang et al. (2018) propose to bridge the source and target embeddings, which brings better attention to the related source and target words. Their method is applicable to any language pairs, providing a tight interaction between the source and target word pairs. However, their method requires additional components and model parameters.

In this work, we aim to enhance the word representations and the interactions between the source and target words, while using even fewer parameters. To this end, we present a language-independent method, which is called shared-private bilingual word embeddings, to share a part of the embeddings of a pair of source and target words that have some common characteristics (i.e. similar words should have similar vectors). Figure 1 illustrates the difference between the standard word embeddings and shared-private word embeddings of NMT. In the proposed method, each source (or target) word is represented by a word embedding that consists of the shared features and the private features. The shared features can also be regarded as the prior alignments connecting the source and target words. The private features allow the words to better learn the monolingual characteristics. Meanwhile, the features shared by the source and target embeddings result in a significant reduction of the number of parameters used for word representations. The experimental results on 6 translation datasets of different scales show that our model with fewer parameters yields consistent improvements over the strong Transformer baselines.

2 Approach

In monolingual vector space, similar words tend to have commonalities in the same dimensions of their word vectors (Mikolov et al., 2013). These commonalities include: (1) a similar degree (value) of the same dimension and (2) a similar positive or negative correlation of the same dimension. Many previous works have noticed this phenomenon and have proposed to use shared vectors to represent similar words in monolingual vector space toward model compression (Li et al., 2016; Zhang et al., 2017b; Li et al., 2018).

Motivated by these works, in NMT, we assume that the source and target words that have similar characteristics should also have similar vectors. Hence, we propose to perform this sharing technique in bilingual vector space. More precisely, we share the features (dimensions) between the paired source and target embeddings (vectors). However, in contrast to the previous studies, we also model the private features of the word embedding to preserve the private characteristics of words for source and target languages. The private

features allow the words to better learn the monolingual characteristics. Meanwhile, we also propose to adopt different sharing mechanisms among the word pairs, which will be described in the following sections.

In the Transformer architecture, the shared features between the source and target embeddings always contribute to the calculation of the attention weight.[2] This results in paying more attention strength on the pair of related words. With the help of residual connections, the high-level representations can also benefit from the shared features of the topmost embedding layers. Both qualitative and quantitative analyses show the effectiveness on the translation tasks.

2.1 Shared-Private Bilingual Word Embeddings

Standard NMT jointly learns to translate and align, which has achieved remarkable results (Bahdanau et al., 2015). In NMT, the intention is to identify the translation relationships between the source and target words. To simplify the model, we propose to divide the relationships into three main categories between a pair of source and target words: (1) words with similar lexical meaning (abbreviated as lm), (2) words with same word form (abbreviated as wf), and (3) unrelated words (abbreviated as ur). Figure 2 shows some examples of these different relationship categories. The number of the shared features of the word embeddings is decided by their relationships.

Before presenting the pairing process in detail, we first introduce the constraints to the proposed method for convenience:

- Each source word is only allowed to share the features with a single target word, and vice versa.[3]
- Each source word preferentially shares features with the target word that has similar lexical meaning, followed by the word with same word form, and then unrelated words.

2.1.1 Words with Similar Lexical Meaning

As shown in Figure 2(a), the English word "Long" and the German word "Lange", which have similar meaning, tend to share more common features of their embeddings. In our model, the source and target words with alignment links are regarded as parallel words that are the translation of each other. According to the word frequency, each source word x is paired with a target aligned word $\hat{y}$ that has the highest alignment probability among the candidates, and is computed as follows:

$$\hat{y} = \underset{y \in a(x)}{\arg\max} \log A(y|x) \quad (1)$$

where $a(\cdot)$ denotes the set of aligned candidates. It is worth noting the target words that have been paired with the source words cannot be used as candidates. $A(\cdot|\cdot)$ denotes the alignment probability. These can be obtained by either the intrinsic attention mechanism (Bahdanau et al., 2015) or unsupervised word aligner (Dyer et al., 2013).

2.1.2 Words with Same Word Form

As shown in Figure 2(b), the sub-word "Ju@@" simultaneously exists in English and German sentences. This kind of word tends to share a medium number of features of the word embeddings. Most of the time, the source and target words with the same word form also share similar lexical meaning. This category of words generally includes Arabic numbers, punctuations, named entities, cognates and loanwords. However, there are some bilingual homographs where the words in the source and target languages look the same but have completely different meanings. For example, the German word "Gift" means "Poison" in English. That is the reason we propose to first pair the words with similar lexical meaning instead of those words with same word forms. This might be the potential limitation of the three-way WT method (Press and Wolf, 2017), where words with the same word form indiscriminately share the same word embedding.

2.1.3 Unrelated Words

We regard source and target words that cannot be paired with each other as unrelated words. Figure 2(c) shows an example of a pair of unrelated words. This category is mainly composed of low-frequency words, such as misspelled words, special characters, and foreign words. In standard NMT, the embeddings of low-frequency words are usually inadequately trained, resulting in a poor word representation. These words are often treated as noises and they are generally ignored

[2] Based on the dot-product attention mechanism, the attention weight between the source and target embeddings is the sum of the dot-product of their features.

[3] We investigate the effect of synonym in the experiment section.

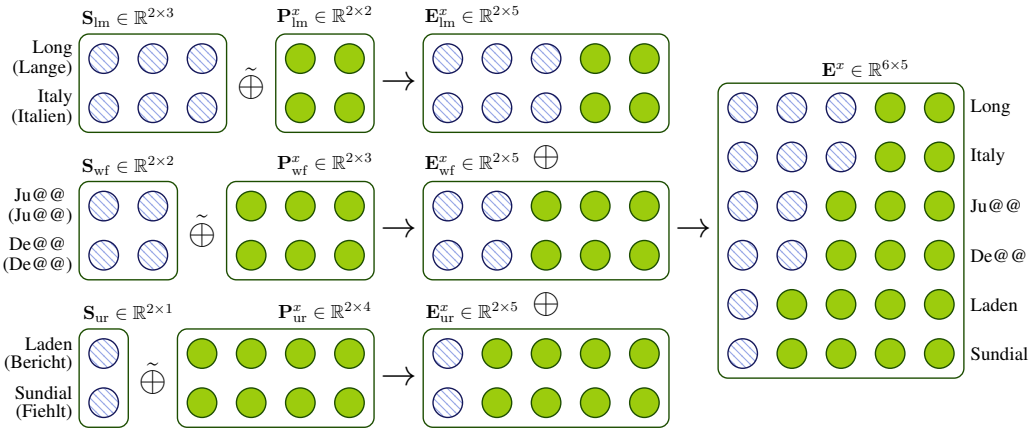


Figure 3: The example of assembling the source word embedding matrix. The words in parentheses denote the paired words sharing features with them.

by the NMT systems (Feng et al., 2017). Motivated by the frequency clustering methods proposed by Chen et al. (2016) where they cluster the words with similar frequency for training a hierarchical language model, in this work, we propose to use a small vector to model the possible features that might be shared between the source and target words which are unrelated but having similar word frequencies. In addition, it can be regarded as a way to improve the robustness of learning the embeddings of low-frequency words because of the noisy dimensions (Wang et al., 2018).

2.2 Implementation

Before looking up embedding at each training step, the source and target embedding matrix are assembled by the sub-embedding matrices. As shown in Figure 3, the source embedding $\mathbf{E}^x \in \mathbb{R}^{|V|\times d}$ is computed as follows::

$$\mathbf{E}^x = \mathbf{E}^x_{\text{lm}} \oplus \mathbf{E}^x_{\text{wf}} \oplus \mathbf{E}^x_{\text{ur}} \quad (2)$$

where $\oplus$ is the row concatenation operator. $\mathbf{E}^x_{(\cdot)} \in \mathbb{R}^{|V_{(\cdot)}|\times d}$ represents the word embeddings of the source words belong to different categories, e.g. lm represents the words with similar lexical meaning. $|V_{(\cdot)}|$ denotes the vocabulary size of the corresponding category.

The process of feature sharing is also implemented by matrix concatenation. For example, the embedding matrices of the source words with similar lexical meaning are computed as follows:

$$\mathbf{E}^x_{\text{lm}} = \mathbf{S}_{\text{lm}} \tilde{\oplus} \mathbf{P}^x_{\text{lm}} \quad (3)$$

where $\tilde{\oplus}$ is the column concatenation operator. $\mathbf{S}_{\text{lm}} \in \mathbb{R}^{|V_{\text{lm}}|\times\lambda_{\text{lm}}d}$ represent the word embeddings of the shared features, where λ_{lm} denotes the proportion of the features for sharing in this relationship category. $\mathbf{P}^x_{\text{lm}} \in \mathbb{R}^{|V_{\text{lm}}|\times(1-\lambda_{\text{lm}})d}$ represent the word embeddings of the private features.

Similar to the target word embedding. These matrix concatenation operations, which have low computational complexity, are very cheap to the whole NMT computation process. We also empirically find both the training speed and decoding speed are not influenced with the introduction of the proposed method.

3 Experiments

We carry out our experiments on the small-scale IWSLT'17 {Arabic (Ar), Japanese (Ja), Korean (Ko), Chinese (Zh)}-to-English (En) translation tasks, medium-scale NIST Chinese-English (Zh-En) translation task, and large-scale WMT'14 English-German (En-De) translation task.

For the IWSLT {Ar, Ja, Ko, Zh}-to-En translation tasks, there are respectively 236K, 234K, 227K, and 235K sentence pairs in each training set.[4] The validation set is IWSLT17.TED.tst2014 and the test set is IWSLT17.TED.tst2015. For each language, we learn a BPE model with 16K merge operations (Sennrich et al., 2016b).

For the NIST Zh-En translation task, the training corpus consists of 1.25M sentence pairs with 27.9M Chinese words and 34.5M English words. We use the NIST MT06 dataset as the validation set and the test sets are the NIST MT02, MT03, MT04, MT05, MT08 datasets. To compare with the recent works, the vocabulary size is limited to

[4] https://wit3.fbk.eu/mt.php?release=2017-01-trnted

Architecture	Zh⇒En	Params	Emb.	Red.	Dev.	MT02	MT03	MT04	MT08	All
SMT*	-	-	-	-	34.00	35.81	34.70	37.15	25.28	33.39
RNNsearch*	Vanilla	74.8M	55.8M	0%	35.92	37.88	36.21	38.83	26.30	34.81
	Source bridging	78.5M	55.8M	0%	36.79	38.71	37.24	40.28	27.40	35.91
	Target bridging	76.6M	55.8M	0%	36.69	39.04	37.63	40.41	27.98	36.27
	Direct bridging	78.9M	55.8M	0%	36.97	39.77	38.02	40.83	27.85	36.62
Transformer	Vanilla	90.2M	46.1M	0%	41.37	42.53	40.25	43.58	32.89	40.33
	Direct bridging	90.5M	46.1M	0%	41.67	42.89	41.34	43.56	32.69	40.54
	Decoder WT	74.9M	30.7M	33.4%	41.90	43.02	41.89	43.87	32.62	40.82
	Shared-private	62.8M	18.7M	59.4%	42.57$^{\uparrow}$	43.73$^{\uparrow}$	41.99$^{\uparrow}$	44.53$^{\uparrow}$	33.81$^{\Uparrow}$	41.61$^{\Uparrow}$

Table 1: Results on the NIST Chinese-English translation task. "Params" denotes the number of model parameters. "Emb." represents the number of parameters used for word representation. "Red." represents the reduction rate of the standard size. The results of SMT* and RNNsearch* are reported by Kuang et al. (2018) with the same datasets and vocabulary settings. "↑" indicates the result is significantly better than that of the vanilla Transformer ($p < 0.01$), while "⇑" indicates the result is significantly better than that of all other Transformer models ($p < 0.01$). All significance tests are measured by paired bootstrap resampling (Koehn, 2004).

En⇒De	Params	Emb.	Red.	BLEU
Vanilla	98.7M	54.5M	0%	27.62
Direct bridging	98.9M	54.5M	0%	27.79
Decoder WT	80.4M	36.2M	33.6%	27.51
Three-way WT	63.1M	18.9M	65.3%	27.39
Shared-private	65.0M	20.9M	63.1%	28.06‡

Table 2: Results on the WMT English-German translation task. "‡" indicates the result is significantly better than the vanilla Transformer model ($p < 0.05$).

	Model	Emb.	Red.	BLEU
Ar⇒ En	Vanilla	23.6M	0%	28.36
	Shared-private	11.8M	50%	29.71$^{\uparrow}$
Ja⇒ En	Vanilla	25.6M	0%	10.94
	Shared-private	13.3M	48.0%	12.35$^{\uparrow}$
Ko⇒ En	Vanilla	25.1M	0%	16.48
	Shared-private	13.2M	47.4%	17.84$^{\uparrow}$
Zh⇒ En	Vanilla	27.4M	0%	19.36
	Shared-private	13.8M	49.6%	21.00$^{\uparrow}$

Table 3: Results on the IWSLT {Ar, Ja, Ko, Zh}-to-En translation tasks. These distant language pairs belonging to 5 different language families and written in 5 different alphabets."↑" indicates the result is significantly better than that of the vanilla Transformer ($p < 0.01$).

30K for both languages, covering 97.7% Chinese words and 99.3% English words, respectively.

For the WMT En-De translation task, the training set contains 4.5M sentence pairs with 107M English words and 113M German words. We use the newstest13 and newstest14 as the validation set and test set, respectively. The joint BPE model is set to 32K merge operations.

3.1 Setup

We implement all of the methods based on Transformer (Vaswani et al., 2017) using the *base* setting with the open-source toolkit *thumt*[5] (Zhang et al., 2017a). There are six encoder and decoder layers in our models, while each layer employs eight parallel attention heads. The dimension of the word embedding and the high-level representation d_{model} is 512, while that of the inner-FFN layer d_{ff} is 2048. The Adam (Kingma and Ba, 2015) optimizer is used to update the model parameters with hyper-parameters β_1= 0.9, β_2 = 0.98, ε = 10^{-8} and a warm-up strategy with $warmup\_steps = 4000$ is adapted to the variable learning rate (Vaswani et al., 2017). The dropout used in the residual connection, attention mechanism, and feed-forward layer is set to 0.1. We employ uniform label smoothing with 0.1 uncertainty.

During the training, each training batch contains nearly 25K source and target tokens. We evaluate the models every 2000 batches via the tokenized BLEU (Papineni et al., 2002) for early stopping. During the testing, we use the best single model for decoding with a beam of 4. The length penalty is tuned on the validation set, which is set to 0.6 for the English-German translation tasks, and 1.0 for others.

We compare our proposed methods with the following related works:

- **Direct bridging** (Kuang et al., 2018): this method minimizes the word embedding loss between the transformations of the target words and their aligned source words by adding an auxiliary objective function.

- **Decoder WT** (Press and Wolf, 2017): this method uses an embedding matrix to repre-

[5] https://github.com/thumt/THUMT

Zh-En	λ_{lm}	λ_{wf}	λ_{ur}	Emb.	BLEU
Vanilla	-	-	-	46.1M	41.37
Decoder WT	0	0	0	30.7M	41.90
Shared-private	0.5	0.7	0.9	21.2M	41.98
	0.5	0.5	0.5	23.0M	42.26
	0.9	0.7	0	21.0M	42.27
	1	1	1	15.3M	42.36
	0.9	0.7	0.5	18.7M	42.57

Table 4: Performance of models using different sharing coefficients on the validation set of the NIST Chinese-English translation task.

sent the target input embedding and target output embedding.

- **Three-way WT** (Press and Wolf, 2017): this method is an extension of the decoder WT method that the source embedding and the two target embeddings are represented by one embedding matrix. This method cannot be applied to the language pairs with different alphabets, e.g. Zh-En.

For the proposed model, we use an unsupervised word aligner *fast-align*[6] (Dyer et al., 2013) to pair source and target words that have similar lexical meaning. We set the threshold of alignment probability to 0.05, i.e. only those words with an alignment probability over 0.05 can be paired as the words having similar lexical meaning. The sharing coefficient $\boldsymbol{\lambda} = (\lambda_{\text{lm}}, \lambda_{\text{wf}}, \lambda_{\text{wf}})$ is set to (0.9,0.7,0.5), which is tuned on both the NIST Chinese-Enlgish task and the WMT English-German task.

3.2 Main Results

Table 1 reports the results on the NIST Chinese-English test sets. It is observed that the Transformer models significantly outperform SMT and RNNsearch models. Therefore, we decide to implement all of our experiments based on Transformer architecture. The direct bridging model can further improve the translation quality of the Transformer baseline. The decoder WT model improves the translation quality while reducing the number of parameters for the word representation. This improved performance happens because there are fewer model parameters, which prevents over-fitting (Press and Wolf, 2017). Finally, the performance is further improved by the proposed method while using even fewer parameters than other models.

[6] https://github.com/clab/fast_align

$A(\cdot\|\cdot)$	Lexical	Form	Unrelated	Emb.	BLEU
0.5	4,869	309	24,822	22.0M	42.35
0.1	15,103	23	14,874	20.0M	42.53
0.05	21,172	11	8,817	18.7M	42.57

Table 5: Effects on different alignment thresholds used for pairing the words with similar lexical meaning on the validation set of the NIST Chinese-English translation task.

Similar observations are obtained on the English-German translation task, as shown in Table 2. The improvement of the direct bridging model is reduced with the introduction of sub-word units since the attention distribution of the high-level representations becomes more confused. Although the two WT methods use fewer parameters, their translation quality degrades. We believe that sub-word NMT needs the well-trained embeddings to distinguish the homographs of sub-words. In the proposed method, both the source and target embeddings benefit from the shared features, which leads to better word representations. Hence, it improves the quality of translation and also reduces the number of parameters.

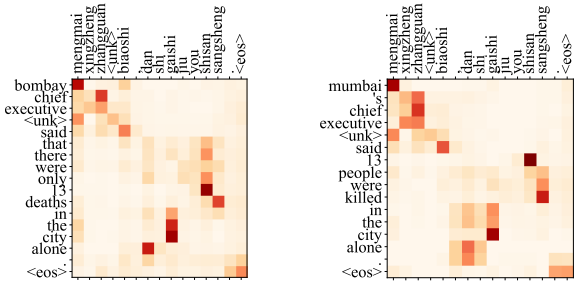
Table 3 shows the results on the small-scale IWSLT translation tasks. We observe that the proposed method stays consistently better than the vanilla model on these distant language pairs. Although the Three-way WT method has been sufficiently validated on similar translation pairs at low-resource settings (Sennrich et al., 2016a), it is not applicable to these distant language pairs. Instead, the proposed method is language-independent, making the WT methods more widely used.

3.3 Effect on Sharing Coefficients

The coefficient $\boldsymbol{\lambda} = (\lambda_{\text{lm}}, \lambda_{\text{wf}}, \lambda_{\text{ur}})$ controls the proportion of the shared features. As shown in Table 4, the decoder WT model can be seen as a kind of shared-private method where *zero* features are shared between the source and target word embeddings. For the proposed method, $\boldsymbol{\lambda} = (0.5, 0.5, 0.5)$ and $\boldsymbol{\lambda} = (1, 1, 1)$ are, respectively, used for sharing half and all features between the embeddings of all categories of words. This allows the model to significantly reduce the number of parameters and also improve the translation quality. For comparison purpose, we also consider sharing a large part of the features among the unrelated words by setting s_3 to 0.9, i.e. $\boldsymbol{\lambda} = (0.5, 0.7, 0.9)$. We argue that it is hard for

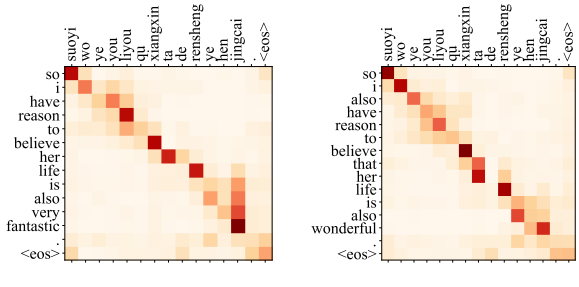
1	Source	mengmai xingzheng zhangguan bazhake biaoshi , dan shi gaishi jiu you shisan **sangsheng** .
	Reference	mumbai municipal commissioner phatak claimed that 13 people were **killed** in the city alone .
	Vanilla	bombay chief executive said that there were only 13 deaths in the city alone .
	Direct bridging	bombay 's chief executive , said there were 13 dead in the city alone .
	Decoder WT	chief executive of bombay , said that thirteen people had died in the city alone .
	Shared-private	mumbai 's chief executive said 13 people were **killed** in the city alone .
2	Source	suoyi wo **ye** you liyou qu xiangxin ta de rensheng **ye** hen jingcai .
	Reference	thus , i **also** have reason to believe that her life is **also** very wonderful .
	Vanilla	so i have reason to believe her life is **also** very fantastic .
	Direct bridging	so i had reason to believe her life was **also** brilliant .
	Decoder WT	so , i have reasons to believe that she has a wonderful life .
	Shared-private	so i **also** have reason to believe that her life is **also** wonderful .

Table 6: Translation examples on MT08 test set. The first and second examples show the accuracy and adequacy of the proposed method, respectively. The **bold** words in each example are paired and will be discussed in the text.



(a) Vanilla (b) Shared-private

Figure 4: Long-distance reordering illustrated by the attention maps. The attention weights learned by the proposed shared-private model is more concentrated than that of the vanilla model.



(a) Vanilla (b) Shared-private

Figure 5: Word omission problem illustrated by the attention maps. In the vanilla model, the third source word "ye" is not translated, while our shared-private model adequately translates it to give a better translation result.

the model to learn an appropriate bilingual vector space in such a sharing setting.

Finally, we propose to share more features between the more similar words by using $s_1 = 0.9$ and reduce the weight on the unrelated words, which is $\boldsymbol{\lambda} = (0.9, 0.7, 0.5)$. This strikes the right balance between the translation quality and the number of model parameters. To investigate whether to share the features between unrelated words or not, we further conduct an experiment with the setting $\boldsymbol{\lambda} = (0.9, 0.7, 0)$. The result confirms our assumption that a small number of shared features between unrelated words with similar word frequency achieve better model performance.

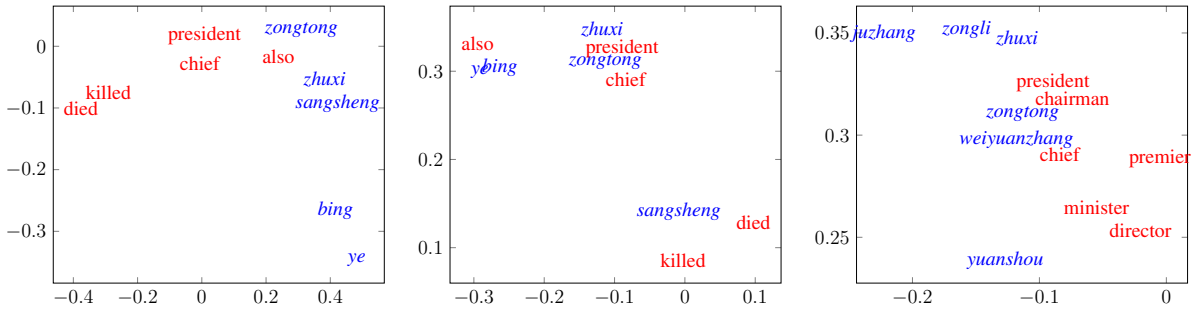
3.4 Effect on Alignment Quality

Table 5 shows the performance of different word alignment thresholds. In the first row, we only pair the words whose alignment probability $A(y|x)$ is above the threshold of 0.5 (see Equation 1 for more details). Under this circumstance, 4,869 words are categorized as parallel words that have similar lexical meaning. Based on these observations, we find that the alignment quality is not a key factor affecting the model performance. In contrast, pairing as many as similar words possible helps the model to better learn the bilingual vector space, which improves the translation performance. The following qualitative analyses support these observations either.

3.5 Analysis of the Translation Results

Table 6 shows two translation examples of the NIST Chinese-English translation task. To better understand the translations produced by these two models, we use layer-wise relevance propagation (LRP) (Ding et al., 2017) to produce the attention maps of the selected translations, as shown in Figure 4 and 5.

In the first example, the Chinese word "sangsheng" is a low-frequency word and its ground truth is "killed". It is observed the inadequate representation of "sangsheng" leads to a decline in the translation quality of the vanilla, direct bridging, and decoder WT methods. In our proposed



(a) Vanilla (b) Shared-private (global) (c) Shared-private (local)

Figure 6: Visualization of the 2-dimensional PCA projection of the bilingual word embeddings of the two models. The *blue* words represent the Chinese embeddings while the red words represent the English embeddings. In (a), only the similar monolingual words are clustered together. While in (b) and (c), both the monolingual and bilingual words which have similar meanings are gathered together.

method, a part of the embedding of "sangsheng" is shared with that of "killed". These improved source representations help the model to generate better translations. Furthermore, as shown in Figure 4, we observe that the proposed method has better long-distance reordering ability than the vanilla. We attribute this improvement to the shared features, which provide an alignment guidance for the attention mechanism.

The second example implies that our proposed model is able to improve the adequacy of translation, as illustrated in Figure 5. The Chinese word "ye" (also) appears twice in the source sentence, while only the proposed method can adequately translate both of them to the target word "also". This once again proves that the shared embeddings between the pair words,"ye" and "also" provide the attention model with a strong interaction between the words, leading to a more concentrated attention distribution and effectively alleviating the word omission problem.

3.6 Analysis of the Learned Embeddings

The proposed method has a limitation in that each word can only be paired with one corresponding word. However, *synonym* is a quite common phenomenon in natural language processing tasks. Qualitatively, we use principal component analysis (PCA) to visualize the learned embeddings of the vanilla model and the proposed method, as shown in Figure 6. In the vanilla model, as shown in Figure 6(a), only the similar monolingual embeddings are clustered, such as the English words "died" and "killed", and the Chinese words "zhuxi" (president) and "zongtong" (president). However, in the proposed method, no matter whether the similar source and target words are paired or not, they tend to cluster together; as shown in Figure 6(b) and 6(c). In other words, the proposed method is able to handle the challenge of synonym. For example, both the Chinese words "ye" (paired with "also") and "bing" can be correctly translated to "also" and these three words tend to gather together in the vector space. This is similar to the Chinese word "sangsheng" (paired with "killed") and the English words "died" and "killed". Figure 6(c) shows that the representations of the Chinese and English words which relate to "president" are very close.

4 Related Work

Many previous works focus on improving the word representations of NMT by capturing the fine-grained (character) or coarse-grained (sub-word) *monolingual* characteristics, such as character-based NMT (Costa-Jussà and Fonollosa, 2016; Ling et al., 2015; Cho et al., 2014; Chen et al., 2016), sub-word NMT (Sennrich et al., 2016b; Johnson et al., 2017; Ataman and Federico, 2018), and hybrid NMT (Luong and Manning, 2016). They effectively consider and utilize the morphological information to enhance the word representations. Our work aims to enhance word representations through the *bilingual* features that are cooperatively learned by the source and target words.

Recently, Gu et al. (2018) propose to use the pre-trained target (English) embeddings as a universal representation to improve the representation learning of the source (low-resource) languages.

In our work, both the source and target embeddings can make use of the common representation unit, i.e. the source and target embedding help each other to learn a better representation.

The previously proposed methods have shown the effectiveness of integrating prior word alignments into the attention mechanism (Mi et al., 2016; Liu et al., 2016; Cheng et al., 2016; Feng et al., 2017), leading to more accurate and adequate translation results with the assistance of prior guidance. We provide an alternative that integrates the prior alignments through the sharing of features, which can also leads to a reduction of model parameters.

Kuang et al. (2018) propose to shorten the path length between the related source and target embeddings to enhance the embedding layer. We believe that the shared features can be seem as the ***zero*** distance between the paired word embeddings. Our proposed method also uses several ideas from the three-way WT method (Press and Wolf, 2017). Both of these methods are easy to implement and transparent to different NMT architectures. The main differences are: 1) we share a part of features instead of all features; 2) the words of different relationship categories are allowed to share with differently sized features; and (3) it is adaptable to any language pairs, making the WT methods more widely used.

5 Conclusion

In this work, we propose a novel sharing technique to improve the learning of word embeddings for NMT. Each word embedding is composed of shared and private features. The shared features act as a prior alignment guidance for the attention model to improve the quality of attention. Meanwhile, the private features enable the words to better capture the monolingual characteristics, result in an improvement of the overall translation quality. According to the degree of relevance between a parallel word pair, the word pairs are categorized into three different groups and the number of shared features is different. Our experimental results show that the proposed method outperforms the strong Transformer baselines while using fewer model parameters.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (Nos. 61672555, 61876035, 61732005), the Joint Project of Macao Science and Technology Development Fund and National Natural Science Foundation of China (No. 045/2017/AFJ), the Multi-Year Research Grant from the University of Macau (No. MYRG2017-00087-FST). Yang Liu is supported by the National Key R&D Program of China (No. 2017YFB0202204), National Natural Science Foundation of China (No. 61761166008, No. 61432013), Beijing Advanced Innovation Center for Language Resources (No. TYR17002).

References

Duygu Ataman and Marcello Federico. 2018. Compositional representation of morphologically-rich input for neural machine translation. In *ACL 2018*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Niki Parmar, Mike Schuster, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *ACL 2018*.

Welin Chen, David Grangier, and Michael Auli. 2016. Strategies for training large vocabulary neural language models. In *ACL 2016*.

Yong Cheng, Shiqi Shen, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Agreement-based joint training for bidirectional attention-based neural machine translation. In *IJCAI 2016*.

Kyunghyun Cho, Bart Van Merriënboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP 2014*.

Marta R. Costa-Jussà and José A. R. Fonollosa. 2016. Character-based neural machine translation. In *ACL 2016*.

Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Visualizing and understanding neural machine translation. In *ACL 2017*.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *NAACL-HLT 2013*.

Y Feng, S Zhang, A Zhang, D Wang, and A Abel. 2017. Memory-augmented neural machine translation. In *EMNLP 2017*.

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

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O K Li. 2018. Universal neural machine translation for extremely low resource languages. In *NAACL-HLT 2018*.

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

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *EMNLP 2013*.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR 2015*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP 2004*.

Shaohui Kuang, Junhui Li, António Branco, Weihua Luo, and Deyi Xiong. 2018. Attention focusing for neural machine translation by bridging source and target embeddings. In *ACL 2018*.

Xiang Li, Tao Qin, Jian Yang, and Tie-Yan Liu. 2016. 2-component recurrent neural networks. In *NIPS 2016*.

Zhongliang Li, Raymond Kulhanek, Shaojun Wang, Yunxin Zhao, and Shuang Wu. 2018. Slim embedding layers for recurrent neural language models. In *AAAI 2018*.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. 2015. Character-based neural machine translation. *arXiv*.

Lemao Liu, Masao Utiyama, Andrew M Finch, and Eiichiro Sumita. 2016. Neural Machine Translation with Supervised Attention. In *COLING 2016*.

Minh-Thang Luong and Christopher D Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *ACL 2016*.

Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. 2016. Supervised attentions for neural machine translation. In *EMNLP 2016*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR 2013*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *ACL 2002*.

Ofir Press and Lior Wolf. 2017. Using the Output Embedding to Improve Language Models. In *EACL 2017*.

Rico Sennrich, Birch, Alexandra, Currey, Anna, Germann, Ulrich, Haddow, Barry, Heafield, Kenneth, Barone, Antonio Valerio Miceli, and Williams, Philip. 2017. The university of edinburgh's neural mt systems for wmt17. In *WMT@EMNLP 2017*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *ACL 2016*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *ACL 2016*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS 2014*.

Joseph P Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *ACL 2010*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS 2017*.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. Switchout: An efficient data augmentation algorithm for neural machine translation. In *EMNLP 2018*.

Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huanbo Luan, and Yang Liu. 2017a. Thumt: An open source toolkit for neural machine translation. *arXiv*.

Xiaowei Zhang, Wei Chen, Feng Wang, Shuang Xu, and Bo Xu. 2017b. Towards compact and fast neural machine translation using a combined method. In *EMNLP 2017*.