Improving Pre-Trained Multilingual Models with Vocabulary Expansion

Hai Wang[1]*   Dian Yu[2]   Kai Sun[3]*   Jianshu Chen[2]   Dong Yu[2]
[1]Toyota Technological Institute at Chicago, Chicago, IL, USA
[2]Tencent AI Lab, Bellevue, WA, USA [3]Cornell, Ithaca, NY, USA
haiwang@ttic.edu, ks985@cornell.edu,
{yudian,jianshuchen,dyu}@tencent.com

Abstract

Recently, pre-trained language models have achieved remarkable success in a broad range of natural language processing tasks. However, in multilingual setting, it is extremely resource-consuming to pre-train a deep language model over large-scale corpora for each language. Instead of exhaustively pre-training monolingual language models independently, an alternative solution is to pre-train a powerful multilingual deep language model over large-scale corpora in hundreds of languages. However, the vocabulary size for each language in such a model is relatively small, especially for low-resource languages. This limitation inevitably hinders the performance of these multilingual models on tasks such as sequence labeling, wherein in-depth token-level or sentence-level understanding is essential.

In this paper, inspired by previous methods designed for monolingual settings, we investigate two approaches (i.e., joint mapping and mixture mapping) based on a pre-trained multilingual model BERT for addressing the out-of-vocabulary (OOV) problem on a variety of tasks, including part-of-speech tagging, named entity recognition, machine translation quality estimation, and machine reading comprehension. Experimental results show that using mixture mapping is more promising. To the best of our knowledge, this is the first work that attempts to address and discuss the OOV issue in multilingual settings.

1 Introduction

It has been shown that performance on many natural language processing tasks drops dramatically on held-out data when a significant percentage of words do not appear in the training data,

---

* This work was done when H. W. and K. S. were at Tencent AI Lab, Bellevue, WA.

i.e., out-of-vocabulary (OOV) words (Søgaard and Johannsen, 2012; Madhyastha et al., 2016). A higher OOV rate (i.e., the percentage of the unseen words in the held-out data) may lead to a more severe performance drop (Kaljahi et al., 2015). OOV problems have been addressed in previous works under monolingual settings, through replacing OOV words with their semantically similar in-vocabulary words (Madhyastha et al., 2016; Kolachina et al., 2017) or using character/word information (Kim et al., 2016, 2018; Chen et al., 2018) or subword information like byte pair encoding (BPE) (Sennrich et al., 2016; Stratos, 2017).

Recently, fine-tuning a pre-trained deep language model, such as Generative Pre-Training (GPT) (Radford et al., 2018) and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), has achieved remarkable success on various downstream natural language processing tasks. Instead of pre-training many monolingual models like the existing English GPT, English BERT, and Chinese BERT, a more natural choice is to develop a powerful multilingual model such as the multilingual BERT.

However, all those pre-trained models rely on language modeling, where a common trick is to tie the weights of softmax and word embeddings (Press and Wolf, 2017). Due to the expensive computation of softmax (Yang et al., 2017) and data imbalance across different languages, the vocabulary size for each language in a multilingual model is relatively small compared to the monolingual BERT/GPT models, especially for low-resource languages. Even for a high-resource language like Chinese, its vocabulary size 10k in the multilingual BERT is only half the size of that in the Chinese BERT. Just as in monolingual settings, the OOV problem also hinders the performance of a multilingual model on tasks that are sensitive to token-level or sentence-level information. For ex-

ample, in the POS tagging problem (Table 2), 11 out of 16 languages have significant OOV issues (OOV rate $\geq 5\%$) when using multilingual BERT.

According to previous work (Radford et al., 2018; Devlin et al., 2018), it is time-consuming and resource-intensive to pre-train a deep language model over large-scale corpora. To address the OOV problems, instead of pre-training a deep model with a large vocabulary, we aim at enlarging the vocabulary size when we fine-tune a pre-trained multilingual model on downstream tasks.

We summarize our contributions as follows: (i) We investigate and compare two methods to alleviate the OOV issue. To the best of our knowledge, this is the first attempt to address the OOV problem in multilingual settings. (ii) By using English as an interlingua, we show that bilingual information helps alleviate the OOV issue, especially for low-resource languages. (iii) We conduct extensive experiments on a variety of token-level and sentence-level downstream tasks to examine the strengths and weaknesses of these methods, which may provide key insights into future directions[1].

2 Approach

We use the multilingual BERT as the pre-trained model. We first introduce the pre-training procedure of this model (Section 2.1) and then introduce two methods we investigate to alleviate the OOV issue by expanding the vocabulary (Section 2.2). Note that these approaches are not restricted to BERT but also applicable to other similar models.

2.1 Pre-Trained BERT

Compared to GPT (Radford et al., 2018) and ELMo (Peters et al., 2018), BERT (Devlin et al., 2018) uses a **bidirectional** transformer, whereas GPT pre-trains a left-to-right transformer (Liu et al., 2018); ELMo (Peters et al., 2018) independently trains left-to-right and right-to-left LSTMs (Peters et al., 2017) to generate representations as additional features for end tasks.

In the pre-training stage, Devlin et al. (2018) use two objectives: masked language model (LM) and next sentence prediction (NSP). In masked LM, they randomly mask some input tokens and then predict these masked tokens. Compared to unidirectional LM, masked LM enables representations to fuse the context from both directions. In the

NSP task, given a certain sentence, it aims to predict the next sentence. The purpose of adding the NSP objective is that many downstream tasks such as question answering and language inference require sentence-level understanding, which is not directly captured by LM objectives.

After pre-training on large-scale corpora like Wikipedia and BookCorpus (Zhu et al., 2015), we follow recent work (Radford et al., 2018; Devlin et al., 2018) to fine-tune the pre-trained model on different downstream tasks with minimal architecture adaptation. We show how to adapt BERT to different downstream tasks in Figure 1 (left).

2.2 Vocabulary Expansion

Devlin et al. (2018) pre-train the multilingual BERT on Wikipedia in 102 languages, with a shared vocabulary that contains 110k subwords calculated from the WordPiece model (Wu et al., 2016). If we ignore the shared subwords between languages, on average, each language has a 1.1k vocabulary, which is significantly smaller than that of a monolingual pre-trained model such as GPT (40k). The OOV problem tends to be less serious for languages (e.g., French and Spanish) that belong to the same language family of English. However, this is not always true, especially for morphologically rich languages such as German (Ataman and Federico, 2018; Lample et al., 2018). OOV problem is much more severe in low-resource scenarios, especially when a language (e.g., Japanese and Urdu) uses an entirely different character set from high-resource languages.

We focus on addressing the OOV issue at subword level in multilingual settings. Formally, suppose we have an embedding E_{bert} extracted from the (non-contextualized) embedding layer in the multilingual BERT (i.e., the first layer of BERT). And suppose we have another set of (non-contextualized) sub-word embeddings $\{E_{l_1}, E_{l_2}, \ldots, E_{l_n}\} \cup \{E_{en}\}$, which are pre-trained on large corpora using any standard word embedding toolkit. Specifically, E_{en} represents the pre-trained embedding for English, and E_{l_i} represents the pre-trained embedding for non-English language l_i at the subword level. We denote the vocabulary of E_{bert}, E_{en}, and E_{l_i} by V_{bert}, V_{en}, and V_{l_i}, respectively. For each subword w in V_{bert}, we use $E_{bert}(w)$ to denote the pre-trained embedding of word w in E_{bert}. $E_{l_i}(\cdot)$ and $E_{en}(\cdot)$ are defined in a similar way as $E_{bert}(\cdot)$. For

---

[1] Improved models will be available at `https://github.com/sohuren/multilingul-bert`.
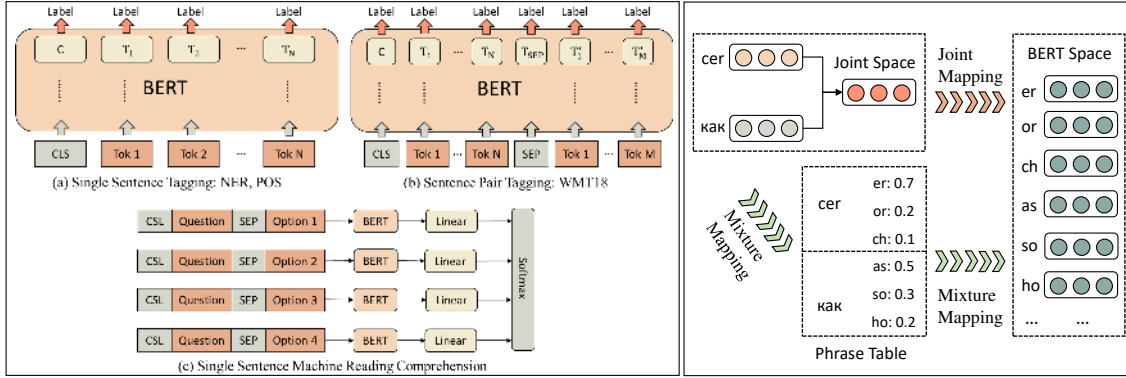
Figure 1: Left: fine-tuning BERT on different kinds of end tasks. Right: illustration of joint and mixture mapping (in this example, during mixture mapping, we represent $e(cer) = 0.7 * e(er) + 0.2 * e(or) + 0.1 * e(ch)$).

each non-English language $l \in \{l_1, l_2, \ldots, l_n\}$, we aim to enrich E_{bert} with more subwords from the vocabulary in E_{l_i} since E_{l_i} contains a larger vocabulary of language l_i compared to E_{bert}.

As there is no previous work to address multilingual OOV issues, inspired by previous solutions designed for monolingual settings, we investigate the following two methods, and all of them can be applied at both word/subword level, though we find subword-level works better (Section 3).

Joint Mapping For each non-English language l, we first construct a joint embedding space E'_l through mapping E_l to E_{en} by an orthogonal mapping matrix B_l (i.e., $E'_l = E_l B_l$). When a bilingual dictionary $f_l : V_l \rightarrow V_{en}$ is available or can be constructed based on the shared common subwords (Section 3.1), we obtain B_l by minimizing:

$$\sum_{w' \in V_l \cap \{w: f_l(w) \in V_{en}\}} \left\| E_l(w') B_l - E_{en}(f_l(w')) \right\|_F^2$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Otherwise, for language pair (e.g., English-Urdu) that meets neither of the above two conditions, we obtain B_l by an unsupervised word alignment method from MUSE (Conneau et al., 2018).

We then map E'_l to E_{bert} by an orthogonal mapping matrix A'_l, which is obtained by minimizing

$$\sum_{w \in f_l(V_l) \cap V_{bert}} \left\| E'_l(w) A'_l - E_{bert}(w) \right\|_F^2$$

We denote this method by M_J in our discussion below, where the subscript J stands for "joint".

Mixture Mapping Following the work of Gu et al. (2018) where they use English as *"universal tokens"* and map all other languages to English

to obtain the subword embeddings, we represent each subword in E'_l (described in joint mapping) as a mixture of English subwords where those English subwords are already in the BERT vocab V_{bert}. This method, denoted by M_M, is also a joint mapping without the need for learning the mapping from E'_l to E_{bert}. Specifically, for each $w \in V_l$, we obtain its embedding $e(w)$ in the BERT embedding space E_{bert} as follows.

$$e(w) = \sum_{u \in \mathcal{T}(w)} p(u \,|\, w) E_{bert}(u)$$

where $\mathcal{T}(w)$ is a set to be defined later, and the mixture coefficient $p(u|w)$ is defined by

$$p(u \,|\, w) = \frac{\exp(\text{CSLS}(E_l(u), E_{en}(w)))}{\sum_{v \in \mathcal{T}(w)} \exp(\text{CSLS}(E_l(v), E_{en}(w)))}$$

where CSLS refers to the distance metric Cross-domain Similarity Local Scaling (Conneau et al., 2018). We select five $v \in V_{en} \cup V_{bert}$ with the highest $\text{CSLS}(E_l(v), E_{en}(w))$ to form set $\mathcal{T}(w)$. In all our experiments, we set the number of nearest neighbors in CSLS to 10. We refer readers to Conneau et al. (2018) for details. Figure 1 (right) illustrates the joint and mixture mapping.

3 Experiment

3.1 Experiment Settings

We obtain the pre-trained embeddings of a specific language by training fastText (Bojanowski et al., 2017) on Wikipedia articles in that language, with context window 5 and negative sampling 5. Before training, we first apply BPE (Sennrich et al., 2016) to tokenize the corpus with subword vocabulary size 50k. For joint mapping method M_J,

we use bilingual dictionaries provided by Conneau et al. (2018). For a language pair where a bilingual dictionary is not easily available, if two languages share a significant number of common subwords (this often happens when two languages belong to the same language family), we construct a bilingual dictionary based on the assumption that identical subwords have the same meaning (Søgaard et al., 2018). We add all unseen subwords from 50k vocabulary to BERT. We define a word as an OOV word once it cannot be represented as a single word. For example, in BERT, the sentence *"Je sens qu' entre ça et les films de médecins et scientifiques"* is represented as *"je sens qu ##' entre [UNK] et les films de [UNK] et scientifiques"*, where **qu'** is an OOV word since it can only be represented by two subword units: **qu** and **##'**, but it is not OOV at subword level; **ça** and **médecins** cannot be represented by any single word or combination of subword units, and thus they are OOV at both word and subword level.

We use the **multilingual BERT** with default parameters in all our experiments, except that we tune the batch size and training epochs. To have a thorough examination about the pros and cons of the explored methods, we conduct experiments on a variety of token-level and sentence-level classification tasks: part of speech (POS) tagging, named entity recognition (NER), machine translation quality estimation, and machine reading comprehension. See more details in each subsection.

3.2 Discussions about Mapping Methods

Previous work typically assumes a linear mapping exists between embedding spaces of different languages if their embeddings are trained using similar techniques (Xing et al., 2015; Madhyastha et al., 2016). However, it is difficult to map embeddings learned with different methods (Søgaard et al., 2018). Considering the differences between BERT and fastText: e.g., the objectives, the way to differentiate between different subwords, and the much deeper architecture of BERT, it is very unlikely that the (non-contextualized) BERT embedding and fastText embedding reside in the same geometric space. Besides, due to that BERT has a relatively smaller vocabulary for each language, when we map a pre-trained vector to its portion in BERT indirectly as previous methods, the supervision signal is relatively weak, especially for low-resource languages. In our experiment,

we find that the accuracy of the mapping from our pre-trained English embedding to multilingual BERT embedding (English portion) is lower than 30%. In contrast, the accuracy of the mapping between two regular English embeddings that are pre-trained using similar methods (e.g., CBOW or SkipGram (Mikolov et al., 2013)) could be above 95% (Conneau et al., 2018).

Besides the **joint mapping** method M_J (Section 2.2), another possible method that could be used for OOV problem in multilingual setting is that, for each language l, we map its pre-trained embedding space E_l to embedding E_{bert} by an orthogonal mapping matrix A_l, which is obtained by minimizing $\sum_{w \in V_l \cap V_{bert}} \|E_l(w)A_l - E_{bert}(w)\|_F^2$. This approach is similar to (Madhyastha et al., 2016), and is referred as **independent mapping** method below. However, we use examples to demonstrate why these kind of methods are less promising. In Table 1, the first two rows are results obtained by mapping our pre-trained English embedding (using fastText) to the (non-contextualized) BERT embedding. In this new unified space, we align words with CSLS metric, and for each subword that appears in English Wikipedia, we seek its closest neighbor in the BERT vocabulary. Ideally, each word should find itself if it exists in the BERT vocabulary. However, this is not always true. For example, although *"however"* exists in the BERT vocabulary, independent mapping fails to find it as its own closest neighbor. Instead, it incorrectly maps it to irrelevant Chinese words "盘" (*"plate"*) and "龙" (*"dragon"*). A similar phenomenon is observed for Chinese. For example, "册" is incorrectly aligned to Chinese words "书" and "卷".

Source Lang	Source	Target	probability
English	however	盘(plate)	0.91
	however	龙(dragon)	0.05
Chinese	册(booklet)	书(book)	0.49
	册(booklet)	卷(volume)	0.46

Table 1: Alignment from Independent Mapping.

Furthermore, our POS tagging experiments (Section 3.3) further show that joint mapping M_J does not improve (or even hurt) the performance of the multilingual BERT. Therefore, we use **mixture mapping** M_M to address and discuss OOV issues in the remaining sections.

	BTS♣	BiLSTM◇	FREQ◇	BERT	BERT$_{oov}$	BERT$_{oovR}$	BERT$_{oovMJ}$	OOV$_w$	OOV$_{sw}$
ar	-	98.23	90.06	53.34	**56.70**	56.57	56.23	89.8	70.6
bg	97.84	98.23	90.06	**98.70**	98.22	94.41	97.21	45.7	1.2
da	95.52	96.16	96.35	**97.16**	96.53	94.15	94.85	38.9	2.8
de	92.87	93.51	93.38	93.58	**93.81**	91.77	93.12	43.2	5.6
es	95.80	95.67	95.74	96.04	**96.92**	95.10	95.77	29.4	6.0
fa	96.82	97.60	97.49	95.62	94.90	94.35	**95.82**	35.6	6.5
fi	95.48	95.74	95.85	87.72	**93.35**	84.75	89.39	64.9	10.4
fr	95.75	96.20	96.11	95.17	**96.59**	94.84	95.24	33.9	10.3
hr	-	96.27	96.82	95.03	**96.49**	89.87	93.48	49.5	8.3
it	97.56	97.90	97.95	**98.22**	98.00	97.63	97.85	30.3	2.3
nl	-	92.82	93.30	**93.89**	92.89	91.30	91.71	35.5	0.3
no	-	98.06	98.03	**97.25**	95.98	94.21	95.83	38.7	4.4
pl	-	97.63	97.62	91.62	**95.95**	87.50	92.56	56.5	13.6
pt	-	97.94	97.90	96.66	**97.63**	95.93	96.90	34.0	8.3
sl	-	96.97	96.84	95.02	**96.91**	89.55	93.97	49.2	7.8
sv	95.57	96.60	96.69	91.23	**96.66**	90.45	91.92	48.2	17.7
average	-	96.60	95.64	92.27	93.60	90.15	92.20	45.2	11.0

Table 2: POS tagging accuracy (%) on the Universal Dependencies v1.2 dataset. BERT$_{oov}$: BERT with method M_M. BERT$_{oovR}$: BERT with randomly picked embedding from BERT. BERT$_{oovMJ}$: BERT with method M_J. OOV$_w$: word-level OOV rate. OOV$_{sw}$: subword-level OOV rate. ♣: Gillick et al. (2016), ◇: Plank et al. (2016).

Approach	Precision	Recall	F1 score
DomainMask (Peng and Dredze, 2017a)	60.8	44.9	51.7
Linear Projection (Peng and Dredze, 2017a)	67.2	41.2	51.1
Updates (Peng and Dredze, 2017b)	-	-	56.1
Updates (Peng and Dredze, 2017b)	-	-	59.0
BERT	56.6	61.7	59.0
BERT$_{oov}$	60.2	62.8	**61.4**
BERT$_{zh}$	63.4	70.8	**66.9**

Table 3: Performance of various models on the test set of Weibo NER. BERT$_{zh}$: Chinese BERT pre-trained over Chinese Wikipedia. We use scripts *conlleval* for evaluation on NER.

3.3 Monolingual Sequence Labeling Tasks

POS Tagging: We use the Universal Dependencies v1.2 data (McDonald et al., 2013). For languages with token segmentation ambiguity, we use the gold segmentation following Plank et al. (2016). We consider languages that have sufficient training data and filter out languages that have unsatisfying embedding alignments with English (accuracy is lower than 20.0% measured by word alignment accuracy or 0.25 by unsupervised metric in MUSE (Conneau et al., 2018)). Finally, we keep 16 languages. We use the original multilingual BERT (without using CRF (Lafferty et al., 2001) on top of it for sequence labeling) to tune hyperparameters on the dev set and use the fixed hyperparameters for the expanded multilingual model. We do not tune the parameters for each model separately. As shown in Table 2, at both the word and subword level, the OOV rate in this dataset is quite high. Mixture mapping improves the accuracy on 10 out of 16 languages,

leading to a 1.97% absolute gain in average. We discuss the influence of alignments in Section 3.6.

Chinese NER: We are also interested in investigating the performance gap between the expanded multilingual model and a monolingual BERT that is pre-trained on a large-scale monolingual corpus. Currently, pre-trained monolingual BERT models are available in English and Chinese. As English has been used as the interlingua, we compare the expanded multilingual BERT and the Chinese BERT on a Chinese NER task, evaluated on the Weibo NER dataset constructed from social media by Peng and Dredze (2015). In the training set, the token-level OOV rate is 2.17%, and the subword-level OOV rate is 0.54%. We tune the hyperparameters of each model based on the development set separately and then use the best hyperparameters of each model for evaluation on the test set.

As shown in Table 3, the expanded model outperforms the multilingual BERT on the Weibo NER dataset. We boost the F1 score from 59.0%

to 61.4%. Compared to the Chinese BERT (66.9%), there still exists a noticeable performance gap. One possible reason could be the grammatical differences between Chinese and English. As BERT uses the language model loss function for pre-training, the pre-trained Chinese BERT could better capture the language-specific information comapred to the multilingual BERT.

3.4 Code-Mixed Sequence Labeling Tasks

As the multilingual BERT is pre-trained over 102 languages, it should be able to handle code-mixed texts. Here we examine its performance and the effectiveness of the expanded model in mixed language scenarios, using two tasks as case studies.

Code-Switch Challenge: We first evaluate on the CALCS-2018 code-switched task (Aguilar et al., 2018), which contains two NER tracks on Twitter social data: mixed English&Spanish (en-es) and mixed Modern Standard Arabic&Egyptian (ar-eg). Compared to traditional NER datasets constructed from news, the dataset contains a significant portion of uncommon tokens like hashtags and abbreviations, making it quite challenging. For example, in the en-es track, the token-level OOV rate is 44.6%, and the subword-level OOV rate is 3.1%; in the ar-eg track, the token-level OOV rate is 64.0%, and the subword-level OOV rate is 6.0%. As shown in Table 4, on ar-eg, we boost the F1 score from 74.7% to 77.3%. However, we do not see similar gains on the en-es dataset, probably because that English and Spanish share a large number of subwords, and adding too many new subwords might prevent the model from utilizing the well pre-trained subwords embedding. See Section 3.6 for more discussions.

Model	en-es			ar-eg		
	Prec	Rec	F1	Prec	Rec	F1
FAIR♣	-	-	62.4	-	-	71.6
IIT♣	-	-	63.8	-	-	-
FAIR◇	-	-	67.7	-	-	**81.4**
BERT	72.7	**63.6**	67.8	73.8	75.6	74.7
BERT$_{oov}$	**74.2**	60.9	66.9	**76.9**	77.8	77.3

Table 4: Accuracy (%) on the code-switch challenge. The top two rows are based on the test set, and the bottom three rows are based on the development set. ♣: results from Aguilar et al. (2018). ◇: results from Wang et al. (2018).

Machine Translation Quality Estimation: All previous experiments are based on well-curated data. Here we evaluate the expanded model on a language generation task, where sometimes the generated sentences are out-of-control. We choose the automatic Machine Translation Quality Estimation task and use Task 2 – word-level quality estimation – in WMT18 (Bojar et al., 2018). Given a source sentence and its translation (i.e., target), this task aims to estimate the translation quality (*"BAD"* or *"OK"*) at each position: e.g., each token in the source and target sentence, each **gap** in the target sentence. We use English to German (en-de) SMT translation. On all three categories, the expanded model consistently outperforms the original multilingual BERT (Table 5)[2].

3.5 Sequence Classification Tasks

Finally, we evaluate the expanded model on sequence classification in a mixed-code setting, where results are less sensitive to unseen words.

Code-Mixed Machine Reading Comprehension: We consider the mixed-language machine reading comprehension task. Since there is no such public available dataset, we construct a new Chinese-English code-mixed machine reading comprehension dataset based on 37,436 unduplicated utterances obtained from the transcriptions of a Chinese and English mixed speech recognition corpus King-ASR-065-1[3]. We generate a multiple-choice machine reading comprehension problem (i.e., a question and four answer options) for each utterance. A question is an utterance with an English text span removed (we randomly pick one if there are multiple English spans) and the correct answer option is the removed English span. Distractors (i.e., wrong answer options) come from the top three closest English text spans, which appear in the corpus, based on the cosine similarity of word embeddings trained on the same corpus. For example, given a question "突然听到 21___，那强劲的鼓点，那一张张脸。" (*"Suddenly I heard 21___, and the powerful drum beats reminded me of the players."*) and four answer options { *"forever"*, *"guns"*, *"jay"*, *"twins"* }, the task is to select the correct answer option *"guns"* (*"21 Guns"* is a song by the American rock band *Green Day*). We split the dataset into training, development, and testing of size 36,636, 400, 400, respectively. An-

---

[2]Our evaluation is based on the development set since the test set is only available to participants, and we could not find the submission teams' performance on the developmenet set.

[3]http://kingline.speechocean.com.

Model	Words in MT			Gaps in MT			Words in SRC		
	F1-BAD	F1-OK	F1-multi	F1-BAD	F1-OK	F1-multi	F1-BAD	F1-OK	F1-multi
Fan et al. (2018)	0.68	0.92	**0.62**	-	-	-	-	-	-
Fan et al. (2018)	0.66	0.92	0.61	0.51	0.98	**0.50**	-	-	-
SHEF-PT♣	0.51	0.85	0.43	0.29	0.96	0.28	0.42	0.80	0.34
BERT	0.58	0.91	0.53	0.47	0.98	0.46	0.48	0.90	0.43
BERT$_{oov}$	0.60	0.91	**0.55**	0.50	0.98	**0.49**	0.49	0.90	**0.44**

Table 5: WMT18 Quality Estimation Task 2 for the en→de SMT dataset. ♣: result from Specia et al. (2018). **MT**: machine translation, e.g., target sentence, **SRC**: source sentence. F1-OK: F1 score for *"OK"* class; F1-BAD: F1 score for *"BAD"* class; F1-multi: multiplication of F1-OK and F1-BAD.

notators manually clean and improve the quality problems by generating more confusing distractors in the dev and testing sets to guarantee that these problems are error-free and challenging.

In this experiment, for each BERT model, we follow its default hyperparameters. As shown in Table 6, the expanded model improves the multilingual BERT (38.1%) by 1.2% in accuracy. Human performance (81.4%) indicates that this is not an easy task even for human readers.

Model	Accuracy	
	Development	Test
BERT$_{en}$	38.2	37.3
BERT	38.7	38.1
BERT$_{oov}$	39.4	**39.3**
BERT$_{zh}$	40.0	**45.0**

Table 6: Accuracy (%) of models on the code-mixed reading comprehension dataset. BERT$_{en}$: pre-trained English BERT. BERT$_{zh}$: pre-trained Chinese BERT.

3.6 Discussions

In this section, we first briefly investigate whether the performance boost indeed comes from the reduction of OOV and then discuss the strengths and weaknesses of the methods we investigate.

First, we argue that it is essential to alleviate the OOV issue in multilingual settings. Taking the POS tagging task as an example, we find that most errors occur at the OOV positions (Table 7 in Section 3.3). In the original BERT, the accuracy of OOV words is much lower than that of non-OOV words, and we significantly boost the accuracy of OOV words with the expanded BERT. All these results indicate that the overall improvement mostly comes from the reduction of OOV.

We also observe that the following factors may influence the performance of the expanded model. **Subwords**: When expanding the vocabulary, it is critical to add only frequent subwords. Currently,

Lang	BERT		BERT$_{oov}$	
	non-OOV	OOV	non-OOV	OOV
fi	98.1	81.3	98.5	90.2
fr	97.0	90.2	97.2	95.6
hr	97.8	91.9	97.7	94.5
pl	98.8	84.6	99.0	93.2
pt	98.8	91.5	98.6	94.8
sl	98.6	91.6	98.7	95.1
sv	97.4	82.9	98.2	94.8
average	98.1	87.7	98.3	94.0

Table 7: POS tagging accuracy (%) for OOV tokens and non-OOV tokens on the Universal Dependencies v1.2 dataset, where the OOV/non-OOV are defined at word level with the original BERT vocabulary.

we add all unseen subwords from the 50k vocabulary (Section 3.1), which may be not an optimal choice. Adding too many subwords may prevent the model from utilizing the information from pretrained subword embedding in BERT, especially when there is a low word-level overlap between the training and test set.

Language: We also find that languages can influence the performance of the vocabulary expansion through the following two aspects: the alignment accuracy and the closeness between a language and English. For languages that are closely related to English such as French and Dutch, it is relatively easy to align their embeddings to English as most subword units are shared (Søgaard et al., 2018; Conneau et al., 2018). In such case, the BERT embedding already contains sufficient information, and therefore adding additional subwords may hurt the performance. On the other hand, for a distant language such as Polish (Slavic family), which shares some subwords with English (Germanic family), adding subwords to BERT brings performance improvements. In the meantime, as Slavic and Germanic are two subdivisions of the Indo-European languages, we find that the embedding alignment methods per-

form reasonably well. For these languages, vocabulary expansion is usually more effective, indicated by POS tagging accuracies for Polish, Portuguese, and Slovenian (Table 2). For more distant languages like Arabic (Semitic family) that use different character sets, it is necessary to add additional subwords. However, as the grammar of such a language is very different from that of English, how to accurately align their embeddings is the main bottleneck.

Task: We see more significant performance gains on NER, POS and MT Quality Estimation, possibly because token-level understanding is more critical for these tasks, therefore alleviating OOV helps more. In comparison, for sequence level classification tasks such as machine reading comprehension (Section 3.5), OOV issue is less severe since the result is based on the entire sentence.

4 Related Work

OOV poses challenges for many tasks (Pinter et al., 2017) such as machine translation (Razmara et al., 2013; Sennrich et al., 2016) and sentiment analysis (Kaewpitakkun et al., 2014). Even for tasks such as machine reading comprehension that are less sensitive to the meanings of each word, OOV still hurts the performance (Chu et al., 2017; Zhang et al., 2018). We now discuss previous methods in two settings.

4.1 Monolingual Setting

Most previous work address the OOV problems in monolingual settings. Before more fine-grained encoding schema such as BPE (Sennrich et al., 2016) is proposed, prior work mainly focused on OOV for token-level representations (Taylor et al., 2011; Kolachina et al., 2017). Besides simply assigning random embeddings to unseen words (Dhingra et al., 2017) or using an unique symbol to replace all these words with a shared embedding (Hermann et al., 2015), a thread of research focuses on refining the OOV representations based on word-level information, such as using similar in-vocabulary words (Luong et al., 2015; Cho et al., 2015; Tafforeau et al., 2015; Li et al., 2016), mapping initial embedding to task-specific embedding (Rothe et al., 2016; Madhyastha et al., 2016), using definitions of OOV words from auxiliary data (Long et al., 2016; Bahdanau et al., 2017), and tracking contexts to build/update representations (Henaff et al., 2016;

Kobayashi et al., 2017; Ji et al., 2017; Zhao et al., 2018).

Meanwhile, there have been efforts in representing words by utilizing character-level (Zhang et al., 2015; Ling et al., 2015a,b; Kim et al., 2016; Gimpel and Livescu, 2016) or subword-level representations (Sennrich et al., 2016; Bojanowski et al., 2017). To leverage the advantages in character and (sub)word level representation, some previous work combine (sub)word- and character-level representations (Santos and Zadrozny, 2014; dos Santos et al., 2015; Yu et al., 2017) or develop hybrid word/subword-character architectures (Chung et al., 2016; Luong and Manning, 2016; Pinter et al., 2017; Bahdanau et al., 2017; Matthews et al., 2018; Li et al., 2018). However, all those approaches assume monolingual setting, which is different from ours.

4.2 Multilingual Setting

Addressing OOV problems in a multilingual setting is relatively under-explored, probably because most multilingual models use separate vocabularies (Jaffe, 2017; Platanios et al., 2018). While there is no direct precedent, previous work show that incorporating multilingual contexts can improve monolingual word embeddings (Zou et al., 2013; Andrew et al., 2013; Faruqui and Dyer, 2014; Lu et al., 2015; Ruder et al., 2017).

Madhyastha and España-Bonet (2017) increase the vocabulary size for statistical machine translation (SMT). Given an OOV source word, they generate a translation list in target language, and integrate this list into SMT system. Although they also generate translation list (similar with us), their approach is still in monolingual setting with SMT. Cotterell and Heigold (2017) train char-level taggers to predict morphological taggings for high/low resource languages jointly, alleviating OOV problems to some extent. In contrast, we focus on dealing with the OOV issue at subword level in the context of pre-trained BERT model.

5 Conclusion

We investigated two methods (i.e., joint mapping and mixture mapping) inspired by monolingual solutions to alleviate the OOV issue in multilingual settings. Experimental results on several benchmarks demonstrate the effectiveness of mixture mapping and the usefulness of bilingual information. To the best of our knowledge, this is

the first work to address and discuss OOV issues at the subword level in multilingual settings. Future work includes: investigating other embedding alignment methods such as Gromov-Wasserstein alignment (Alvarez-Melis and Jaakkola, 2018) upon more languages; investigating approaches to choose the subwords to be added dynamically.

Acknowledgments

References

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Overview of the CALCS 2018 Shared Task: Named Entity Recognition on Code-switched Data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147, Melbourne, Australia.

David Alvarez-Melis and Tommi Jaakkola. 2018. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the EMNLP*, pages 1881–1890, Brussels, Belgium.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *Proceedings of the ICML*, pages 1247–1255, Atlanta, GA.

Duygu Ataman and Marcello Federico. 2018. Compositional representation of morphologically-rich input for neural machine translation. In *Proceedings of the ACL*, pages 305–311, Melbourne, Australia.

Dzmitry Bahdanau, Tom Bosc, Stanisław Jastrzebski, Edward Grefenstette, Pascal Vincent, and Yoshua Bengio. 2017. Learning to compute word embeddings on the fly. *arXiv preprint arXiv:1706.00286*.

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.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the WMT*, pages 272–303, Belgium, Brussels.

Huadong Chen, Shujian Huang, David Chiang, Xinyu Dai, and Jiajun Chen. 2018. Combining character and word information in neural machine translation using a multi-level attention. In *Proceedings of the NAACL-HLT*, pages 1284–1293.

Sébastien Jean Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the ACL-IJCNLP*, pages 1–10, Beijing, China.

Zewei Chu, Hai Wang, Kevin Gimpel, and David McAllester. 2017. Broad context language modeling as reading comprehension. In *Proceedings of the EACL*, pages 52–57, Valencia, Spain.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. *arXiv preprint arXiv:1603.06147*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the ICLR*, Vancouver, Canada.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of the EMNLP*, pages 748–759, Copenhagen, Denmark.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proceedings of the ACL*, pages 1832–1846.

Kai Fan, Bo Li, Fengming Zhou, and Jiayi Wang. 2018. " bilingual expert" can find translation errors. *arXiv preprint arXiv:1807.09433*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the EACL*, pages 462–471, Gothenburg, Sweden.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. Multilingual language processing from bytes. In *Proceedings of NAACL-HLT*, pages 1296–1306.

John Wieting Mohit Bansal Kevin Gimpel and Karen Livescu. 2016. Charagram: Embedding words and sentences via character n-grams.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the NAACL-HLT*, pages 344–354, New Orleans, LA.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2016. Tracking the world state with recurrent entity networks. *arXiv preprint arXiv:1612.03969*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Alan Jaffe. 2017. Generating image descriptions using multilingual data. In *Proceedings of the WMT*, pages 458–464, Copenhagen, Denmark.

Yangfeng Ji, Chenhao Tan, Sebastian Martschat, Yejin Choi, and Noah A Smith. 2017. Dynamic entity representations in neural language models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1830–1839.

Yongyos Kaewpitakkun, Kiyoaki Shirai, and Masnizah Mohd. 2014. Sentiment lexicon interpolation and polarity estimation of objective and out-of-vocabulary words to improve sentiment classification on microblogging. In *Proceedings of the PACLIC*, pages 204–213, Phuket, Thailand.

Rasoul Kaljahi, Jennifer Foster, Johann Roturier, Corentin Ribeyre, Teresa Lynn, and Joseph Le Roux. 2015. Foreebank: Syntactic analysis of customer support forums. In *Proceedings of the EMNLP*, pages 1341–1347.

Yeachan Kim, Kang-Min Kim, Ji-Min Lee, and SangKeun Lee. 2018. Learning to generate word representations using subword information. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2551–2561.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *AAAI*, pages 2741–2749.

Sosuke Kobayashi, Naoaki Okazaki, and Kentaro Inui. 2017. A neural language model for dynamically representing the meanings of unknown words and entities in a discourse. *arXiv preprint arXiv:1709.01679*.

Prasanth Kolachina, Martin Riedl, and Chris Biemann. 2017. Replacing oov words for dependency parsing with distributional semantics. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 11–19.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Bofang Li, Aleksandr Drozd, Tao Liu, and Xiaoyong Du. 2018. Subword-level composition functions for learning word embeddings. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 38–48.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2016. Towards zero unknown word in neural machine translation. In *IJCAI*, pages 2852–2858.

Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015a. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. 2015b. Character-based neural machine translation. *arXiv preprint arXiv:1511.04586*.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by summarizing long sequences. In *Proceedings of the ICLR*, Vancouver, Canada.

Teng Long, Ryan Lowe, Jackie Chi Kit Cheung, and Doina Precup. 2016. Leveraging lexical resources for learning entity embeddings in multi-relational data. In *The 54th Annual Meeting of the Association for Computational Linguistics*, page 112.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256.

Minh-Thang Luong and Christopher D Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1054–1063.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 11–19.

Pranava Swaroop Madhyastha, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Mapping unseen words to task-trained embedding spaces. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 100–110.

325

Pranava Swaroop Madhyastha and Cristina España-Bonet. 2017. Learning bilingual projections of embeddings for vocabulary expansion in machine translation. In *Proceedings of the RepL4NLP*, pages 139–145.

Austin Matthews, Graham Neubig, and Chris Dyer. 2018. Using morphological knowledge in open-vocabulary neural language models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1435–1445.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 92–97.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554.

Nanyun Peng and Mark Dredze. 2017a. Multi-task domain adaptation for sequence tagging. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 91–100.

Nanyun Peng and Mark Dredze. 2017b. Supplementary results for named entity recognition on chinese social media with an updated dataset. Technical report.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the ACL*, pages 1756–1765, Vancouver, Canada.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237.

Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. 2017. Mimicking word embeddings using subword rnns. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 102–112.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 412–418.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the EMNLP*, pages 425–435, Brussels, Belgium.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 157–163.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Majid Razmara, Maryam Siahbani, Reza Haffari, and Anoop Sarkar. 2013. Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. In *Proceedings of the ACL*, pages 1105–1115, Sofia, Bulgaria.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902*.

Cıcero dos Santos, Victor Guimaraes, RJ Niterói, and Rio de Janeiro. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of NEWS 2015 The Fifth Named Entities Workshop*, page 25.

Cicero D Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1818–1826.

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

Anders Søgaard and Anders Johannsen. 2012. Robust learning in random subspaces: Equipping nlp for oov effects. *Proceedings of COLING 2012: Posters*, pages 1171–1180.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the ACL*, pages 778–788, Melbourne, Australia.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón Astudillo, and André FT Martins. 2018. Findings of the wmt 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709.

Karl Stratos. 2017. A sub-character architecture for korean language processing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 721–726.

Jeremie Tafforeau, Thierry Artieres, Benoit Favre, and Frederic Bechet. 2015. Adapting lexical representation and oov handling from written to spoken language with word embedding. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Julia M Taylor, Victor Raskin, and Christian F Hempelmann. 2011. Towards computational guessing of unknown word meanings: The ontological semantic approach. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Changhan Wang, Kyunghyun Cho, and Douwe Kiela. 2018. Code-switched named entity recognition with embedding attention. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 154–158.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the NAACL-HLT*, pages 1006–1011, Denver, CO.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W Cohen. 2017. Breaking the softmax bottleneck: A high-rank rnn language model. *arXiv preprint arXiv:1711.03953*.

Jinxing Yu, Xun Jian, Hao Xin, and Yangqiu Song. 2017. Joint embeddings of chinese words, characters, and fine-grained subcharacter components. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 286–291.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Zhuosheng Zhang, Yafang Huang, and Hai Zhao. 2018. Subword-augmented embedding for cloze reading comprehension. In *Proceedings of the COLING*, pages 1802–1814, Santa Fe, NM.

Yang Zhao, Jiajun Zhang, Zhongjun He, Chengqing Zong, and Hua Wu. 2018. Addressing troublesome words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 391–400.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE ICCV*, pages 19–27, Santiago, Chile.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the EMNLP*, pages 1393–1398, Seattle, WA.