BERT for Monolingual and Cross-Lingual Reverse Dictionary

Hang Yan, Xiaonan Li, Xipeng Qiu[*], Bocao Deng
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University
2005 Songhu Road, Shanghai, China
{hyan19,xnli20,xpqiu}@fudan.edu.cn, dengbocao@gmail.com

Abstract

Reverse dictionary is the task to find the proper target word given the word description. In this paper, we tried to incorporate BERT into this task. However, since BERT is based on the byte-pair-encoding (BPE) subword encoding, it is nontrivial to make BERT generate a word given the description. We propose a simple but effective method to make BERT generate the target word for this specific task. Besides, the cross-lingual reverse dictionary is the task to find the proper target word described in another language. Previous models have to keep two different word embeddings and learn to align these embeddings. Nevertheless, by using the Multilingual BERT (mBERT), we can efficiently conduct the cross-lingual reverse dictionary with one subword embedding, and the alignment between languages is not necessary. More importantly, mBERT can achieve remarkable cross-lingual reverse dictionary performance even without the parallel corpus, which means it can conduct the cross-lingual reverse dictionary with only corresponding monolingual data. Code is publicly available at `https://github.com/yhcc/BertForRD.git`.

1 Introduction

Reverse dictionary (Bilac et al., 2004; Hill et al., 2016) is the task to find the proper target word given the word description. Fig. 1 shows an example of the monolingual and the cross-lingual reverse dictionary. Reverse dictionary should be a useful tool to help writers, translators, and new language learners find a proper word when encountering the tip-of-the-tongue problem (Brown and McNeill, 1966). Moreover, the reverse dictionary can be used for educational evaluation. For example, teachers can ask the students to describe a word, and the correct description should make the reverse dictionary model recall the word.

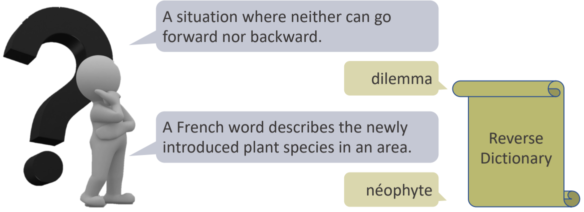
A situation where neither can go forward nor backward.
dilemma
A French word describes the newly introduced plant species in an area.
néophyte
Reverse Dictionary

Figure 1: An example of the monolingual and cross-lingual reverse dictionary.

The core of reverse dictionary is to match a word and its description semantically. Early methods (Bilac et al., 2004; Shaw et al., 2013) firstly extracted the handcrafted features and then used similarity-based approaches to find the target word. However, since these methods are mainly based on the surface form of words, they cannot extract the semantic meaning, resulting in bad performance when evaluated on the human-written search query. Recent methods usually adopt neural networks to encode the description and the candidate words into the same semantic embedding space and return the word which is closest to the description (Hill et al., 2016; Zhang et al., 2019).

Although current neural methods can extract the semantic representations of the descriptions and words, they have three challenging issues: (1) The first issue is the data sparsity. It is hard to learn good embeddings for the low-frequent words; (2) The second issue is polysemy. The previous methods usually use the static word embedding (Mikolov et al., 2013; Pennington et al., 2014), making them struggle to find the target word when the target word is polysemous. Pilehvar (2019) used different word senses to represent a word. Nonetheless, gathering senses for all words is not easy; (3) The third issue is the alignment of cross-lingual word embeddings in the cross-lingual re-

[*]Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4329–4338
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

verse dictionary scenario (Hill et al., 2016; Chen et al., 2019).

In this paper, we leverage the pre-trained masked language model BERT (Devlin et al., 2019) to tackle the above issues. Firstly, since BERT tokenizes the words into subwords with byte-pair-encoding (BPE) (Sennrich et al., 2016b), the common subwords between low-frequent and high-frequent words can alleviate the data sparsity problem. Secondly, BERT can output contextualized representation for a word. Thus the polysemy problem can be much relieved. Thirdly, the mBERT is suitable to tackle the cross-lingual reverse dictionary. Because BERT shares some subwords between different languages, there is no need to align different languages explicitly. Therefore, we formulate the reverse dictionary task into the masked language model framework and use BERT to deal with the reverse dictionary task in monolingual and cross-lingual scenarios. Besides, our proposed framework can also tackle the cross-lingual reverse dictionary task without the parallel (aligned) corpus.

Our contributions can be summarized as follows:

1. We propose a simple but effective solution to incorporate BERT into the reverse dictionary task. In the method, the target word is predicted according to masked language model predictions. With BERT, we achieve significant improvement for the monolingual reverse dictionary task.

2. By leveraging the Multilingual BERT (mBERT), we extend our methods into the cross-lingual reverse dictionary task, mBERT can not only avoid the explicit alignment between different language embeddings, but also achieve good performance.

3. We propose the unaligned cross-lingual reverse dictionary scenario and achieve encouraging performance only with monolingual reverse dictionary data. As far as we know, this is the first time the unaligned cross-lingual reverse dictionary is inspected.

2 Related Work

The reverse dictionary task has been investigated in several previous academic studies. Bilac et al. (2004) proposed using the information retrieval techniques to solve this task, and they first built a database based on available dictionaries. When a query came in, the system would find the closest definition in the database, then return the corresponding word. Different similarity metrics can be used to calculate the distance. Shaw et al. (2013) enhanced the retrieval system with WordNet (Miller, 1995). Hill et al. (2016) was the first to apply RNN into the reverse dictionary task, making the model free of handcrafted features. After encoding the definition into a dense vector, this vector is used to find its nearest neighbor word. This model formulation has been adopted in several papers (Pilehvar, 2019; Chen et al., 2019; Zhang et al., 2019; Morinaga and Yamaguchi, 2018; Hedderich et al., 2019), their difference lies in usage of different resources. Kartsaklis et al. (2018); Thorat and Choudhari (2016) used WordNet to form graphs to tackle the reverse dictionary task.

The construction of the bilingual reverse dictionary has been studied in (Gollins and Sanderson, 2001; Lam and Kalita, 2013). Lam and Kalita (2013) relied on the availability of lexical resources, such as WordNet, to build a bilingual reverse dictionary. Chen et al. (2019) built several bilingual reverse dictionaries based on the Wiktionary[1], but this kind of online data cannot ensure the data's quality. Building a bilingual reverse dictionary is not an easy task, and it will be even harder for low-resource language. Other than the low-quality problem, the vast number of language pairs is also a big obstacle, since if there are N languages, they will form N^2 pairs. However, by the unaligned cross-lingual reverse dictionary, we can not only exploit the high-quality monolingual dictionaries, but also avoid the preparation of N^2 language pairs.

Unsupervised machine translation is highly correlated with the unaligned cross-lingual reverse dictionary (Lample et al., 2018a; Conneau and Lample, 2019; Sennrich et al., 2016a). However, the unaligned cross-lingual reverse dictionary task differs from the unsupervised machine translation at least in two aspects. Firstly, the target for the cross-lingual reverse dictionary and machine translation is a word and a sentence, respectively. Secondly, theoretically, the translated sentence and the original sentence should contain the same information. Nevertheless, in the cross-lingual reverse dictionary task, on the one hand, the target word might contain more senses when it is polysemous. On the other hand, a description can correspond to several

[1] https://www.wiktionary.org/

similar terms. The polysemy also makes the unsupervised word alignment hard to solve this task (Lample et al., 2018b).

Last but not least, the pre-trained language model BERT has been extensively exploited in the Natural Language Processing (NLP) community since its introduction (Devlin et al., 2019; Conneau and Lample, 2019). Owing to BERT's ability to extract contextualized information, BERT has been successfully utilized to enhance various tasks substantially, such as the aspect-based sentiment analysis task (Sun et al., 2019), summarization (Zhong et al., 2019), named entity recognition (Yan et al., 2019; Li et al., 2020) and Chinese dependency parsing (Yan et al., 2020). However, most works used BERT as an encoder, and less work uses BERT to do generation (Wang and Cho, 2019; Conneau and Lample, 2019). Wang and Cho (2019) showed that BERT is a Markov random field language model. Therefore, sentences can be sampled from BERT. Conneau and Lample (2019) used pre-trained BERT to initialize the unsupervised machine training model an achieve good performance. Different from these work, although a word might contain several subwords, we use a simple but effective method to make BERT generate the word ranking list with only one forward pass.

3 Methodology

The reverse dictionary task is to find the target word w given its definition $d = [w_1, w_2, \ldots, w_n]$, where d and w can be in the same language or different languages. In this section, we first introduce BERT, then present the method we used to incorporate BERT into the reverse dictionary task.

3.1 BERT

BERT is a pre-trained model proposed in (Devlin et al., 2019). BERT contains several Transformer Encoder layers. BERT can be formulated as follows

$$\hat{h^l} = \mathrm{LN}(h^{l-1} + \mathrm{MHAtt}(h^{l-1})), \tag{1}$$

$$h^l = \mathrm{LN}(\hat{h^l} + \mathrm{FFN}(\hat{h^l})), \tag{2}$$

where h^0 is the BERT input, for each token, it is the sum of its token embedding, position embedding, and segment embedding; LN is the layer normalization layer; MHAtt is the multi-head self-attention; FFN contains three layers, the first one is a linear projection layer, then an activation layer, then another linear projection layer; l is the depth of the layer, the total number of layers in BERT is 12 or 24.

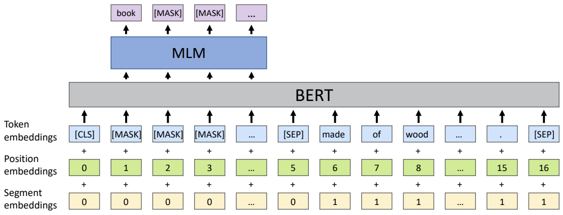

Figure 2: The model structure for the monolingual and cross-lingual reverse dictionary. The "[MASK]" in the input is the placeholder where BERT needs to predict. Placeholders concatenate with the word definition before sending it into BERT. Postprocessing is required to convert the prediction for "[MASK]"s into the word ranking list.

Two tasks were used to pre-train BERT. The first is to replace some tokens with the "[MASK]" symbol, BERT has to recover this masked token from outputs of the last layer. The second one is the next sentence prediction. For two continuous sentences, 50% of the time the second sentence will be replaced with other sentences, BERT has to figure out whether the input sequence is continuous based on the output vector of the "[CLS]" token. Another noticeable fact about BERT is that, instead of directly using the word, it used BPE subword (Sennrich et al., 2016b) to represent tokens.Therefore, one word may be split into several tokens. Next, we will show how we make BERT generate the word ranking list.

3.2 BERT for Monolingual Reverse Dictionary

The model structure is shown in Fig. 2. The input sequence x has the form "[CLS] + [MASK] $*$ k + [SEP] + [subword sequence of the definition d] + [SEP]". We want BERT to recover the target word w from the k "[MASK]" tokens based on the definition d. We first utilize BERT to predict the masks as in its pre-training task. It can be formulated as

$$S_{subword} = \mathrm{MLM}(H_k^L), \tag{3}$$

where $H_k^L \in \mathbb{R}^{k \times d_{model}}$ is the hidden states for the k masked tokens in the last layer, MLM is the pre-trained masked language model, $S_{subword} \in \mathbb{R}^{k \times |V|}$ is the subword score distribution for the k positions, $|V|$ is the number of subword tokens. Although we can make BERT directly predict word by using a word embedding, it will suffer from at least two problems: the first one is that it cannot take advantage of common subwords between words, such as prefixes and postfixes; the second one is that predicting word is inconsistent with the pre-trained tasks.

After achieving $S_{subword}$, we need to convert them back to word scores. However, there are $|V|^k$ kinds of subword combinations, which makes it intractable to represent words by crossing subwords. Another method is to generate subword one-by-one (Wang and Cho, 2019; Conneau and Lample, 2019), it is not suitable for this task, since this task needs to return a ranking list of words, but the generation can only offer limited answers. Nevertheless, for this specific task, the number of possible target words is fixed since the number of unique words in one language's dictionary is limited. Hence, instead of combining the subword sequence into different words, we can only care for the subword sequence, which can form a valid word.

Specifically, for a given language, we first list all its valid words and find the subword sequence for each word. For a word w with the subword sequence $[b_1, ..., b_k]$, its score is calculated by

$$S_{word} = \sum_{i=1}^{k} S^i_{subword}[b_i], \tag{4}$$

where $S_{word} \in \mathbb{R}$ is the score for the word w, $S^i_{subword} \in \mathbb{R}^{|V|}$ is the subword score distribution in the ith position, $S^i_{subword}[b_i]$ is gathering the b_ith element in $S^i_{subword}$. However, not all words can be decomposed to k subword tokens. If a word has subword tokens less than k, we pad it with "[MASK]", while our method cannot handle words with more than k subword tokens. By this method, each word can get a score. Therefore we can directly use the cross-entropy loss to finetune the model,

$$L_w = -\sum_{i=1}^{N} w^{(i)} \text{log\_softmax}(S^{(i)}_{word}), \tag{5}$$

where N is the total number of samples, w is the target word. When ranking, words are sorted by their scores.

3.3 BERT for Cross-lingual Reverse Dictionary

The model structure used in this setting is as depicted in Fig. 2. The only difference between this setting and the monolingual scenario is the pre-trained model used. This setting uses the mBERT model. mBERT has the same structure as BERT, but it was trained on 104 languages. Therefore its token embedding contains subwords in different languages.

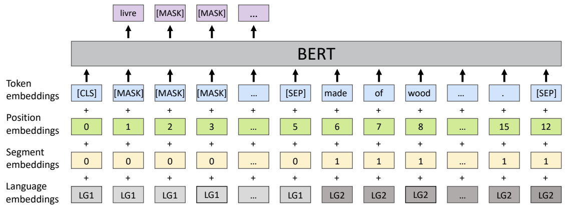

Figure 3: The model structure for the unaligned cross-lingual reverse dictionary. We add a randomly initialized language embedding to distinguish languages. Since we only have monolingual training data, "LG1" and "LG2" are of the same value in the training phase, but different in the evaluation phase.

3.4 BERT for Unaligned Cross-lingual Reverse Dictionary

The model used for this setting is as depicted in Fig. 3. Compared with the BERT model, we add an extra learnable language embedding in the bottom, and the language embedding has the same dimension as the other embeddings. Except for the randomly initialized language embedding, the model is initialized with the pre-trained mBERT.

Instead of using the MLM to get $S_{subword}$, we use the following equation to get $S_{subword}$

$$S_{subword} = H^L_k Emb^T_{token}, \tag{6}$$

where $Emb_{token} \in \mathbb{R}^{|V| \times d_{model}}$ is the subword token embeddings. We found this formulation will lead to better performance than using the MLM, and we assume this is because the training data only contains monolingual data, thus it will be hard for the model to predict tokens in another language when evaluation, while if the Emb_{token} is used, the model can utilize the similarity between subwords to make reasonable predictions. After getting $S_{subword}$, we use Eq.4 to get the scores for each word, and different languages have different word lists, the loss is calculated by

$$L_w = -\sum_{j=1}^{M} \sum_{i=1}^{N_j} w^{(i)}_j \text{log\_softmax}(S^{(i)}_{word_j}), \tag{7}$$

where M is the number languages, N_k is the number of samples for language j, $w^{(i)}_j$ is the target

Language	Word	Type	Train	Dev	Seen	Unseen	Description	Question
English	50.5K	Def	675.7K	75.9K	500	500	200	-
		Word	45.0K	5.0K	500	500	200	-
Chinese	58.5K	Def	78.3K	8.7K	2.1K	2.0K	200	272
		Word	54.0K	6.1K	1.4K	1.4K	200	272

Table 1: Dataset statistics for the monolingual reverse dictionary. The row "Def" and "Word" are the number of definition and distinct words in the split, respectively.

word in language j, $S^{(i)}_{word_j}$ is the score distribution for words in language j. When getting the ranking list for a language, we only calculate word scores for that language.

4 Experimental Setup

4.1 Dataset

For the monolingual reverse dictionary, we tested our methods in the English dataset and Chinese dataset released by (Hill et al., 2016) and (Zhang et al., 2019), respectively. Hill et al. (2016) built this dataset by extracting words and definitions from five electronic dictionaries and Wikipedia. Zhang et al. (2019) used the authoritative Modern Chinese Dictionary to build the Chinese reverse dictionary. There are four different test sets: (1) **Seen** definition set, words and their definitions are seen during the training phase; (2) **Unseen** definition set, none of the word's definitions have been seen during the training phase, but they might occur in other words' definition; (3) **Description** definition set, the description and its corresponding word are given by human. Methods rely on word matching may not perform well in this setting (Hill et al., 2016); (4) **Question** definition set, this dataset is only in Chinese, it contains 272 definitions appeared in Chinese exams. The detailed dataset statistics are shown in Table 1.

For the cross-lingual and unaligned cross-lingual reverse dictionary, we use the dataset released in (Chen et al., 2019). This dataset includes four bilingual reverse dictionaries: English↔French, English↔Spanish. Besides, this dataset includes English, French, and Spanish monolingual reverse dictionary data. The test set for this dataset is four bilingual reverse dictionaries: En↔Fr and En↔Es. For the cross-lingual reverse dictionary, we use the paired bilingual reverse dictionary data to train our model; for the unaligned cross-lingual reverse dictionary, we use the three monolingual reverse dictionary data to train our model. And for both settings, we report results on the test sets of the four bilingual reverse dictionary. The detailed dataset statistics are shown in Table 2.

Scenario	Language	Word	Type	Train	Dev	Test
Monolingual	En	117.4K	Def	228.2K	500	501
			Word	117.3K	499	501
	Fr	52.4K	Def	104.4K	500	501
			Word	52.2K	496	501
	Es	22.5K	Def	47.6K	500	501
			Word	22.4K	493	501
Bilingual	En-Fr	45.6K	Def	49.7K	500	501
			Word	15.6K	493	488
	Fr-En	44.5K	Def	58.1K	500	501
			Word	16.8K	487	486
	En-Es	45.6K	Def	20.2K	500	501
			Word	7.9K	484	495
	Es-En	35.8K	Def	55.9K	500	501
			Word	15.9K	489	487

Table 2: Dataset statistics for the cross-lingual and unaligned cross-lingual reverse dictionary. The upper block is the monolingual data used to train the unaligned cross-lingual reverse dictionary. The lower block is the cross-lingual reverse dictionary data. Both scenarios were evaluated in the test set in the lower part. "En-fr" means the target word is in English, the definition is in French.

4.2 Evaluation Metrics

For the English and Chinese monolingual reverse dictionary, we report three metrics: the median rank of target words (Median Rank, lower better, lowerest is 0), the ratio that target words appear in top 1/10/100 (Acc@1/10/100, higher better, ranges from 0 to 1), and the variance of the rank of the correct target word (Rank Variance, lower better), these results are also reported in (Hill et al., 2016; Zhang et al., 2019). For the cross-lingual and unaligned cross-lingual reverse dictionary, we report the Acc@1/10, and the mean reciprocal rank (MRR, higher is better, ranges from 0 to 1), these results are also reported in (Chen et al., 2019).

4.3 Hyper-parameter Settings

The English BERT and Multilingual BERT (mBERT) are from (Devlin et al., 2019), the Chinese BERT is from (Cui et al., 2019). Since RoBERTa has the same model structure as BERT, we also report the performance with the English RoBERTa from (Liu et al., 2019) and the Chinese RoBERTa from (Cui et al., 2019) for the monolingual reverse dictionary. Both RoBERTa and BERT are the base version, and we use the uncased English BERT and cased mBERT. For all models, we find the hyper-parameters based on the Acc@10 in the development sets, the models with the best development set performance are evaluated on the test set. The data and detailed hyper-parameters for each setting will be released within the code [2]. We choose $k = 4$ for Chinese, and $k = 5$ for other languages, k is determined by at least 99% of the target words in the training set are included.

5 Experimental Results

5.1 Monolingual Reverse Dictionary

Results for the English and Chinese monolingual reverse dictionary have been shown in Table 3 and Table 4, respectively. "OneLook" in Table 3 is the most used commercial reverse dictionary system, it indexed over 1061 dictionaries, even included online dictionaries, such as Wikipedia and WordNet (Miller, 1995). Therefore, its result in the unseen definition test set is ignored. "SuperSense", "RDWECI", "MS-LSTM" and "Mul-Channel" are from (Pilehvar, 2019; Morinaga and Yamaguchi, 2018; Kartsaklis et al., 2018; Zhang et al., 2019), respectively. From Table 3, RoBERTa achieves state-of-the-art performance on the human description test set. And owing to bigger models, in the seen definition test set, compared with the "Mul-channel", BERT and RoBERTa enhance the performance significantly. Although the MS-LSTM (Kartsaklis et al., 2018) performs remarkably in the seen test sets, it fails to generalize to unseen and description test sets. Besides, "RDWECI", "SuperSense", "Mul-channel" in Table 3 all used external knowledge, such as WordNet, Part-of-Speech tags. Combining BERT and structured knowledge should further improve the performance in all test sets, we leave it for further work.

Model	Seen			Unseen			Description		
OneLook*	0	.66/.94/.95	200	-	-	-	5.5	**.33**/.54/.76	332
RDWECI	121	.06/.20/.44	420	170	.05/.19/.43	420	16	.14/.41/.74	306
SuperSense	378	.03/.15/.36	462	465	.02/.11/.31	454	115	.03/.15/.47	396
MS-LSTM*	**0**	**.92/.98/.99**	65	276	.03/.14/.37	426	1000	.01/.04/.18	404
Mul-Channel	16	.20/.44/.71	310	54	.09/.29/.58	358	2	.32/.64/.88	203
BERT	**0**	.57/.86/.92	240	**18**	**.20/.46/.64**	418	1	.36/.77/.94	94
RoBERTa	**0**	.57/.84/.92	228	37	.10/.36/.60	405	**1**	**.43/.85/.96**	46

Table 3: Results on the English reverse dictionary datasets. In each cell, the values are the "Median Rank", "Acc@1/10/100" and "Rank Variance". * results are from (Zhang et al., 2019) . BERT and RoBERTa achieve a significant performance boost in both the description test set and the unseen test set.

Table 4 presents the results for the Chinese reverse dictionary. For the seen definition setting, BERT and RoBERTa substantially improve the performance. Apart from the good performance in seen definitions, BERT and RoBERTa perform well in the human description test set, which depicts their capability to capture human's meaning.

Model	Seen			Unseen			Description			Question		
BOW*	59	.08/.28	403	65	.08/.28	411	40	.07/.30	357	42	.10/.28	362
RDWECI*	56	.09/.31	423	83	.08/.28	436	32	.09/.32	376	45	.12/.32	384
Mul-Channel*	1	.49/.78	220	10	.18/.49	310	5	.24/.56	260	0	.50/.73	223
BERT	**0**	**.88/.93**	201	**5**	**.27/.56**	360	**3**	**.34/.67**	260	**0**	**.57**/.70	325
RoBERTa	**0**	**.88/.93**	200	**5**	**.28/.56**	350	**3**	**.33/.65**	230	**0**	**.59/.74**	310

Table 4: Results on the Chinese reverse dictionary datasets. In each cell, the values are the "Median Rank", "Acc@1/10" and "Rank Variance". * results are from (Zhang et al., 2019). Our proposed methods enhance the performance in all test sets substantially.

5.2 Cross-lingual Reverse Dictionary

In this section, we will present the results for the cross-lingual reverse dictionary. The performance comparison is shown in Table 5, mBERT substantially enhances the performance in four test sets. The contrast between "mBERT" and "mBERT-joint" shows that jointly train the reverse dictionary in different language pairs can improve the performance.

5.3 Unaligned Cross-lingual Reverse Dictionary

In this section, we present the results of the unaligned bilingual and cross-lingual reverse dictionary. Models are trained on several monolingual reverse dictionary data, but they will be evaluated on bilingual reverse dictionary data. Take the "En-Fr" as an example, models are trained on English

Model	En-Fr		Fr-En		En-Es		Es-En	
ATT*	.39/.47	.41	.40/.50	.43	.52/.59	.53	.60/.68	.63
mBERT	.88/.90	.89	.88/.90	.89	.79/.81	.80	.88/.90	.89
ATT-joint*	.64/.69	.65	.68/.75	.71	.69/.73	.70	.79/.83	.80
mBERT-joint	**.90/.94**	.92	**.90/.93**	.91	**.83/.88**	.85	**.93/.95**	.93

Table 5: Results for the cross-lingual reverse dictionary. In each cell, the values are "Acc@1/10" and "MRR". * results are from (Chen et al., 2019). "En-Fr" means the target word is in English, while the description is in French. The "ATT" and "mBERT" used the bilingual corpus to train the model. The "ATT-joint" and "mBERT-joint" are trained on four bilingual reverse dictionary corpus simultaneously.

[2] www.github.com/xxx/will_release

Model	En-Fr		Fr-En		En-Es		Es-En	
ATT-joint*	.64/.69	.65	.68/.75	.71	.69/.73	.70	.79/.83	.80
BERT-joint	.90/.94	.92	.90/.93	.91	.83/.88	.85	.93/.95	.93
BERT-Match	.35/.41	-	.20/.25	-	.23/.26	-	.17/.21	-
BERT-Trans	.46/.55	-	.42/.51	-	.44/.49	-	.29/.38	-
BERT-Unaligned	.70/**.80**	.74	.55/.66	.59	.52/**.68**	.58	**.41/.59**	.48
BERT-joint-Unaligned	**.71/.80**	.74	**.56/.67**	.60	**.54/.68**	.59	**.41/.59**	.47

Table 6: Results for the unaligned cross-lingual reverse dictionary. In each cell, the values are "Acc@1/10" and "MRR". * is from (Chen et al., 2019). "En-Fr" means the target word is in English, while the definition is in French. Models in the lower block do not use aligned data. While models in the upper block use aligned data to train the model.

definitions to English words, French definitions to French words, while in the evaluation phase, the model is asked to recall an English word given the French description or vice versa.

Since previous models do not consider this setting, we make a baseline by firstly getting words with the same language as the definition through a monolingual reverse dictionary model, then using the word translation or aligned word vectors to recall words in another language. Take "En-Fr" for instance, we first recall the top 10 French words with the French definition, then each French word is translated into an English word by either translations or word vectors.

Models listed in Table 6 are as follows: (1) **mBERT-Match** uses aligned word vectors (Lample et al., 2018b) to recall the target words in another language; (2) **mBERT-Trans** uses the translation API[3]; (3) **mBERT-Unaligned** uses two monolingual reverse dictionary corpus to train one model. Therefore, the results of "En-Fr" and "Fr-En" in Table 6 are from the same model; (4) **mBERT-joint-Unaligned** is trained on all monolingual corpus.

As shown in the Table 6, the "mBERT-Unaligned" and "mBERT-joint-Unaligned" perform much better than the "mBERT-Match" and "mBERT-Trans". Therefore, it is meaningful to explore the unaligned reverse dictionary scenario. As we will show in Section 6.4, the translation method might fail to recall the target words when the word is polysemous.

From Table 6, we can see that jointly training three monolingual reverse dictionary tasks do not help to recall cross-lingual words. Therefore, how to utilize different languages to enhance the performance of the unaligned reverse dictionary is an unsolved problem. Besides, compared with the top block of Table 6, the performance of the unaligned models lags much behind. Hence, there is a lot of room for unaligned performance improvement.

6 Analysis

6.1 Performance for Number of Senses

Following (Zhang et al., 2019), we evaluate the accuracy of words with a different number of senses through WordNet(Miller, 1995). The results are shown in Fig. 4. BERT and RoBERTa significantly improve the accuracy of words with single and multiple senses, which means they can alleviate the polysemous issue.

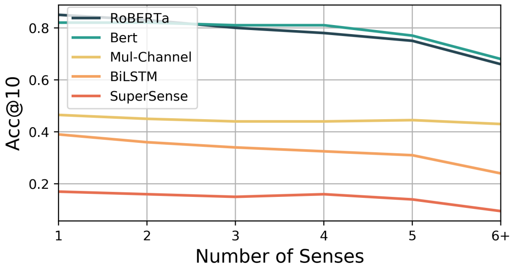

Figure 4: The Acc@10 for English words with a different number of senses.

6.2 Performance for Different Number of Subword

Since BERT decomposes words into subwords, we want to investigate whether the number of subwords has an impact on performance. We evaluate the English development set, results are shown in Fig. 5. The model achieves the best accuracy in English words with one subword and Chinese words with two subwords. This might be caused by the fact that most English words and Chinese words have one subword and two subwords, respectively.

6.3 Unseen Definition in Unaligned Cross-lingual Reverse Dictionary

In this section, for the target words presented in bilingual test sets, we gradually remove their definitions from the monolingual training corpus. The performance changing curve is depicted in Table 6. As a reminder, the test sets need to recall target words in another language, while the deleted word and definition are in the same language. Since the number of removed samples is less than 2% of the monolingual corpus, the performance decay cannot

[3] fanyi.baidu.com

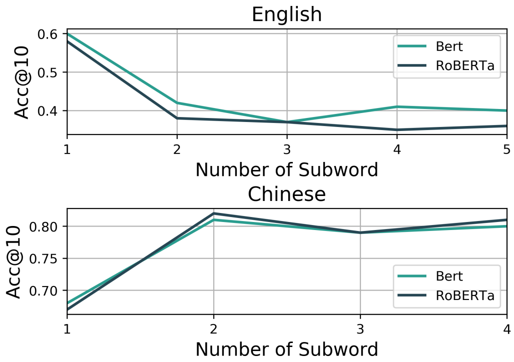

Figure 5: The Acc@10 for words with a different number of subwords.

be totally ascribed to the reducing data. Based on Table 6, for the unaligned reverse dictionary task, we can enhance the cross-lingual word retrieval by including more monolingual word definitions.

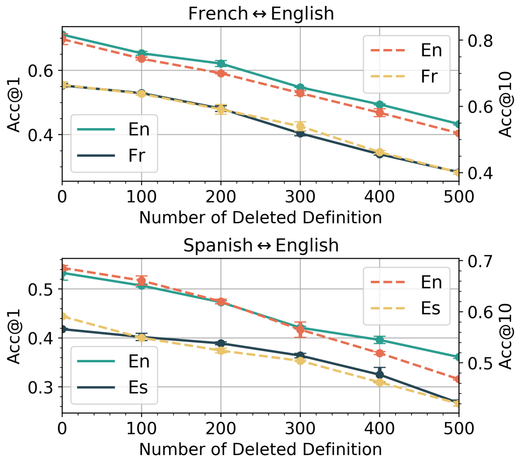

Figure 6: The performance for the unaligned reverse dictionary with the increment of deleted definitions in monolingual data. The dense and dotted lines are Acc@1, Acc@10, respectively. Although the deleted definition and word are in the same language, deleting them harms the performance of cross-lingual word retrieval.

6.4 Case Study

For the monolingual scenario, we present an example in Table 7 to show that decomposing words into subwords helps to recall related words. Table 8 shows the comparison between "mBERT-Trans" and "mBERT-joint-Unaligned".

Definition	someone who studies secret code systems in order to obtain secret information
Mul-Channel	cryptographer cryptologist spymaster snoop
BERT	cryptanalyst codebreaker cryptographer coder
RoBERTa	codebreaker cryptanalyst cryptographer snooper

Table 7: A Monolingual case displays the advantage of using subwords. In each row is the model's top recalled words; the underlined word is the target word. The predicted words by BERT or RoBERTa is either related to "someone" (corresponding to the "-analyst" or "er") or "code/secret" (correspoding to "code-" or "crypt-").

Definition	El punto que esta a mitad del camino entre dos extremos. (The point that is halfway between two ends)			
Spanish	centro	mitad	medio	punta
Trans.	core	middle	middle	tip
Unaligned	center	centre	middle	mid
Definition	Pièce où l ' on prépare et fait cuire les aliments (Room where food is prepared and cooked)			
French	cuisine	restaurant	pièce	cuire
Trans.	cookery	restaurant	room	cook
Unaligned	kitchen	cook	office	restaurant

Table 8: Unaligned reverse dictionary results by translation and the proposed unaligned reverse dictionary model. The target word is underlined, the "Trans." row is the word translation results. The Spanish "centro" in the upper block also has the meaning "center", but without context, it gives the wrong translation, and the French word "cuisine" in the lower block makes the same error.

7 Conclusion

In this paper, we formulate the reverse dictionary task under the masked language model framework and use BERT to predict the target word. Since BERT decomposes words into subwords, the score of the target word is the sum of the scores of its constituent subwords. With the incorporation of BERT, our method achieves state-of-the-art performances for both the monolingual and cross-lingual reverse dictionary tasks. Besides, we propose a new cross-lingual reverse dictionary task without aligned data. Our proposed framework can perform the cross-lingual reverse dictionary while being trained on monolingual corpora only. Although the performance of unaligned BERT is superior to the translation and word vector alignment method, it still lags behind the supervised aligned reverse dictionary model. Therefore, future work should be conducted to enhance performance on the unaligned reverse dictionary.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. We also thank the developers of fastNLP[4], Yunfan Shao and Yining Zheng, to develop this handy natural language processing package. This work was supported by the National Natural Science Foundation of China (No. 61751201, 62022027 and 61976056), Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01) and ZJLab.

References

Slaven Bilac, Wataru Watanabe, Taiichi Hashimoto, Takenobu Tokunaga, and Hozumi Tanaka. 2004. Dictionary search based on the target word description. In *Proceedings of NLP*.

Roger Brown and David McNeill. 1966. The "tip of the tongue" phenomenon. *Journal of verbal learning and verbal behavior*, 5(4):325–337.

Muhao Chen, Yingtao Tian, Haochen Chen, Kai-Wei Chang, Steven Skiena, and Carlo Zaniolo. 2019. Learning to represent bilingual dictionaries. In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pages 152–162. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *NeurIPS*, pages 7057–7067.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese BERT. *CoRR*, abs/1906.08101.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Tim Gollins and Mark Sanderson. 2001. Improving cross language information retrieval with triangulated translation. In *SIGIR*, pages 90–95.

Michael A Hedderich, Andrew Yates, Dietrich Klakow, and Gerard de Melo. 2019. Using multi-sense vector embeddings for reverse dictionaries. In *Proceedings of IWCS*.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *TACL*, 4:17–30.

Dimitri Kartsaklis, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Mapping text to knowledge graph entities using multi-sense LSTMs. In *Proceedings of EMNLP*.

Khang Nhut Lam and Jugal Kumar Kalita. 2013. Creating reverse bilingual dictionaries. In *HLT-NAACL*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *ICLR*.

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

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. FLAT: chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6836–6842. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the Acm*, 38(11):39–41.

Yuya Morinaga and Kazunori Yamaguchi. 2018. Improvement of reverse dictionary by tuning word vectors and category inference. In *Proceedings of ICIST*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Mohammad Taher Pilehvar. 2019. On the importance of distinguishing word meaning representations: A case study on reverse dictionary mapping. In *NAACL-HLT*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *ACL*.

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

Ryan Shaw, Anindya Datta, Debra E. VanderMeer, and Kaushik Dutta. 2013. Building a scalable database-driven reverse dictionary. *TKDE*, 25:528–540.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *NAACL-HLT*, pages 380–385.

Sushrut Thorat and Varad Choudhari. 2016. Implementing a reverse dictionary, based on word definitions, using a node-graph architecture. In *COLING*.

[4] https://github.com/fastnlp/fastNLP

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a markov random field language model. *CoRR*, abs/1902.04094.

Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. 2019. TENER: adapting transformer encoder for named entity recognition. *CoRR*, abs/1911.04474.

Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. A graph-based model for joint chinese word segmentation and dependency parsing. *Trans. Assoc. Comput. Linguistics*, 8:78–92.

Lei Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2019. Multi-channel reverse dictionary model. *CoRR*, abs/1912.08441.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019. Searching for effective neural extractive summarization: What works and what's next. In *ACL*, pages 1049–1058.