Toward Better Loanword Identification in Uyghur Using Cross-lingual Word Embeddings

Chenggang Mi[†‡], Yating Yang[†‡], Lei Wang[†‡], Xi Zhou[†‡], Tonghai Jiang[†‡]
[†]Xinjiang Technical Institute of Physics & Chemistry
of Chinese Academy of Sciences, China
[‡]Key Laboratory of Speech Language Information Processing of Xinjiang, China
{micg,yangyt,wanglei,zhouxi,jth}@ms.xjb.ac.cn

Abstract

To enrich vocabulary of low resource settings, we proposed a novel method which identify loanwords in monolingual corpora. More specifically, we first use cross-lingual word embeddings as the core feature to generate semantically related candidates based on comparable corpora and a small bilingual lexicon; then, a log-linear model which combines several shallow features such as pronunciation similarity and hybrid language model features to predict the final results. In this paper, we use Uyghur as the receipt language and try to detect loanwords in four donor languages: Arabic, Chinese, Persian and Russian. We conduct two groups of experiments to evaluate the effectiveness of our proposed approach: loanword identification and OOV translation in four language pairs and eight translation directions (Uyghur-Arabic, Arabic-Uyghur, Uyghur-Chinese, Chinese-Uyghur, Uyghur-Persian, Persian-Uyghur, Uyghur-Russian, and Russian-Uyghur). Experimental results on loanword identification show that our method outperforms other baseline models significantly. Neural machine translation models integrating results of loanword identification experiments achieve the best results on OOV translation(with 0.5-0.9 BLEU improvements).

1 Introduction

Almost every natural language processing (NLP) task suffers from data sparseness. This situation is even worse for low resource languages in cross lingual tasks, such as neural machine translation(NMT).Lexical borrowing happens in every language. It is a phenomenon of cross-linguistic influence. A loanword is a word adopted from one language (the donor language) and incorporated into another language (the recipient language) without translation. One reason we choose Uyghur as the example in this study is that Uyghur has been influenced by many languages, both oriental and western languages; another reason is that Uyghur language is low-resource and lack of research in NLP field. If loanwords in a resource poor language can be identified effectively, we can use the loanword and its corresponding donor word to extend the bilingual dictionary (Tsvetkov et al., 2015).

Some researchers take loanword identification as a string similarity problem. They have applied several commonly used methods to detect loanwords (Mi et al., 2014). However, these methods are usually rule-based, so these methods are difficult to deal with ambiguity. Machine learning based models have been proposed to overcome the shortcomings that exist in previous work. Lack of annotated corpora weakens the performance of these methods (Mi et al., 2016). Although some studies combine rule based and machine learning based methods to overcome data sparsity to a certain extent, challenges still exist due to semantic issues.

A loanword usually shares the same/similar semantic space with its corresponding source word in donor language. Word embeddings were first proposed to learn representations where words that have the same meaning have a similar representation (Mikolov et al., 2013). Cross-lingual word embedding is one way to learn word embeddings using information from multiple languages (Klementiev et al., 2012). This inspired us to introduce the cross-lingual embedding into the loanword identification. Additionally,

Proceedings of the 27th International Conference on Computational Linguistics, pages 3027–3037
Santa Fe, New Mexico, USA, August 20-26, 2018.

the loanword often has the similar pronunciation with its source word, the string similarity and character-level language model of receipt and donor languages can reflect this characteristic.

This paper proposes a novel loanword identification model based on both shallow features and semantically-related features. We first use a cross-lingual word embedding based model to generate the loanword candidates; then, an optimized loanword identification model based on shallow features such as pronunciation similarity and hybrid language model is built to output the final results. We conduct experiments on loanword identification in Uyghur and OOV translation in several low-resource language pairs separately.[1] Experimental results show that our proposed approach outperforms other baseline models significantly. The output of our loanword identification model can improve translation results by at least 0.5 BLEU points.

2 Related Work

Our study is mainly inspired by following two fields in NLP: loanword identification in NLP and cross-lingual word embedding.

2.1 Loanword Identification in NLP

Loanwords were mainly studied by linguists (Hoffer et al., 2005; Peperkamp, 2004; Kang et al., 2014; Shinohara, 2015) in the early period; there have been relatively few papers in loanwords research in natural language processing (NLP). (Tsvetkov and Dyer, 2016; Tsvetkov and Dyer, 2016a)proposed a morph-phonological transformation model, where features are based on optimality theory; experiments proved that with a few training examples, this model can obtain good performance at predicting donor forms from borrowed forms. (Tsvetkov et al., 2015a) suggest an approach that uses the lexical borrowing as a model in SMT framework to translate OOV words in a low-resource language. (Mi et al., 2014; Mi et al., 2016) use shallow features such as string similarity to detect loanwords in Uyghur.

2.2 Cross-lingual Word Embedding

Recent advances in cross-lingual word embedding have shown its effectiveness in several cross-lingual NLP tasks (Gouws et al., 2015; Ruder et al., 2017; Duong et al., 2017; Ammar et al., 2016), especially in some under-resourced situations (Adams et al., 2017; Fang and Cohn, 2017; Wei and Deng, 2017). Cross-lingual word embedding has the capacity to learn high quality embeddings even in the absence of bilingual corpora by exploiting bilingual lexicons. This is very useful when creation of high quality lexicons on some low-resource languages.

Previous work mainly focused on one specific language or some context free features. These methods are difficult to adopt in situations such as when the receipt and donor language belong to very different language families or low-resource settings. In this study, we propose a novel loanword identification approach, which is not only based on some shallow features, but also cross-lingual word embeddings.

3 Background

3.1 Loanwords in Uyghur

Uyghur is an official language of the Xinjiang Uyghur Autonomous Region in Western China, and is spoken by 10 to 25 million people. Uyghur is an agglutinative language. Due to the different kinds of language contact through the history of the Xinjiang region, Uyghur has also adopted many loanwords (most of them are nouns) from Persian. Additionally, many words of Arabic origin have also entered the language directly through Islamic literature after the introduction of Islam. Among all languages Uyghur borrowed, Russian and Chinese have the greatest influence in recent years. Due to the globalization, a number of loanwords of European origin have also reached Uyghur through Russia's influence over the region. Loanwords in Uyghur not only include named entities such as person and location names, but also some daily used words (Kontovas, 2008; Sugar, 2017) (Figure 1).

[1] Actually there are two methods to overcome OOV problems in neural machine translation in low-resource settings: 1) use different translation granularities (Nguyen and Chiang, 2017; Sennrich et al., 2016); 2) extend the vocabulary. Our experiments on OOV translation belong to the second.

Donor language	Source word	Uyghur word	English
Persian	افسوس	ئەپسۇس	pity
	گوشت	گۆش	meat
Arabic	ساعة	سائەت	hour
Russian	велосипед	ۋېلىسىپىت	bicycle
	доктор	دوختۇر	doctor
	поезд	پويىز	train
	область	ئوبلاست	oblast
	телевизор	تېلېۋىزور	television
Chinese	凉粉	لەڭپۇڭ	agar-agar jelly
	豆腐	دۇفۇ	tofu

Figure 1: Examples of loanwords in Uyghur.

3.2 Challenges in Loanword Identification

Intuitively, loanwords can be simply detected according to pronunciation similarity. However, the words may have several changes to adopt the recipient language:1) Changes in meaning: words are occasionally changed with a different meaning than that in the donor language; 2) Changes in spelling: although words taken into different recipient languages are sometimes spelled as in the donor language, sometimes borrowed words retain their original (or near-original) pronunciation, but undergo a spelling change to represent the orthography of the recipient language; 3) Changes in pronunciation: in cases where a new loanword has a very unusual sound, the pronunciation of the word can be radically changed.

All above challenges shown that the lexical borrowing is very complicated, so it is difficult to achieve expected results without further semantic information, especially some cross-lingual features.

4 Methodology

In this paper, we propose a novel method to identify loanwords in Uyghur. Inspired by state-of-the-art achievements of word embedding techniques in cross-lingual NLP tasks, we first train a cross-lingual embedding model based on comparable corpora (Uyghur-one donor language) and a small bilingual lexicon; then, given a Uyghur word, we obtain its most sematic related source words (and its semantic distance) in donor language based on cross-lingual embedding model; we set a threshold ρ for each donor language, if a Uyghur word has a semantic distance (the first source word of donor language) lower than this threshold, it will be removed. With the candidate list, we use a log-linear model with pronunciation similarity and hybrid language model features to predict the final results (Figure 2).

4.1 Problem Description

Assume we have a low resource language named $Lang_A$ (receipt language) and a high resource language named $Lang_B$ (donor language). The goal of loanword identification is to find out words w_i in the $Lang_A$ corpus which is originally borrowed from $Lang_B$. In this study, we mainly focus on loanwords that share the same or similar meaning with its corresponding word in the donor language.

4.2 Loanword Candidate List Generation

In this section, we first introduce cross-lingual word embedding model based on comparable data, which is used in our proposed approach as an important feature to represent the semantic relationships between words in donor language and receipt language; then, we present the generation of loanword candidate list.

Cross-lingual Word Embedding Model.

Cross-lingual embedding models learn cross-lingual representations of words in a joint embedding space. They enable knowledge transfer between languages, which is very important between resource-rich language and resource-poor language. Many previous studies on cross-lingual word embedding critically

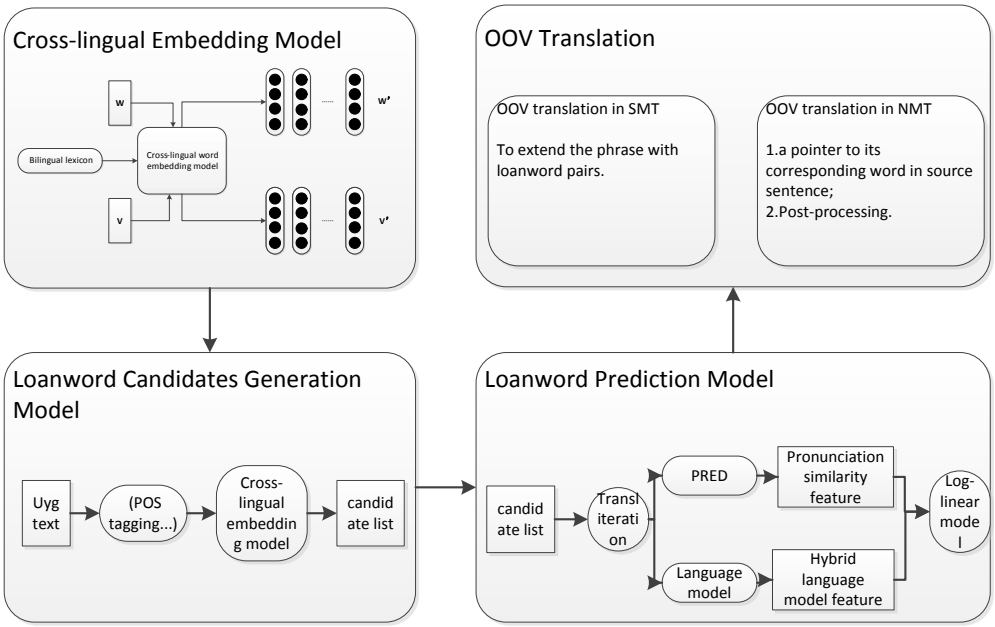

Figure 2: Framework of our proposed model.

require large sentence-aligned parallel data or dictionaries to induce bilingual word emebddings that closely aligned over languages in the same semantic space. However, these bilingual resources are very expensive for low-resource languages like Uyghur. Therefore, in this study we follow the work described by (Vulić and Moens, 2015) to obtain cross-lingual word embeddings with comparable data.

Let $W = \{w_1, w_2, ..., w_{|w|}\}$ be the vocabulary of a language $lang_A$ with $|W|$ words, and $\mathbf{W} \in \mathbb{R}^{|W| \times l}$ be the corresponding word embeddings of length l. Let $V = \{v_1, v_2, ..., v_{|V|}\}$ be the vocabulary of another language $lang_B$ with $|V|$ words, and $\mathbf{V} \in \mathbb{R}^{|V| \times m}$ the corresponding word embeddings of length m. Let d_u(Uyghur) and d_f (donor language) denote a pair of comparable documents with length in words p and q ($p > q$). The comparable data based cross-lingual embeddings (**CdbCLE**) method merges d_u and d_f as one single pseudo-bilingual document exploiting a deterministic strategy based on length ratio of two documents $R = \lfloor \frac{p}{q} \rfloor$ firstly. We sequentially pick word from d_f and put it into the merged document as the Rth word. Then, we train a skip-gram model on merged documents to generate word vectors for all words in $\mathbf{W} \cup \mathbf{V}$. The most important step is to instantiate CdbCLE algorithm. We use the similar objective function with word2vec skip-gram based on pseudo-lingual document described above.

The CdbCLE method can be described as:

$$C(\mathbf{W}, \mathbf{V}) = - \sum_{s \in \mathbf{W} \cup \mathbf{V}} \sum_{t \in NBR(s)} \log P(t|s) \tag{1}$$

where $NBR(s)$ is the context of s in pseudo-lingual document, $P(ts) \propto \exp(t^T s)$. Note that $t, s \in \mathbf{W} \cup \mathbf{V}$.

Candidate List Generation

In this part, we treat the loanword candidate list generation in Uyghur as a procedure to find out the most semantic similar source words in donor language corpora with current Uyghur word.

According to the CdbCLE algorithm described above, to identify loanwords in Uyghur, we must ensure that the current Uyghur word should share the same meaning with the corresponding source word in donor language. This can be indicates that the loanword and its corresponding word in the donor language share the same word embedding space.

To simplify this step and filter some unrelated words in Uyghur, we annotate the Uyghur texts with a part-of-speech tagger firstly. The goal of part-of-speech tagging is to assign to each word in a sentence

its morph-syntactic category. Since loanwords in Uyghur are mainly nouns, we can filter Uyghur word list with POS tags.

We use the CdbCLE algorithm to train a cross-lingual embedding model, the source part is the Uyghur language corpus, and the target parts are the donor language corpora (Chinese, Arabic, Persian and Russian). When testing, the pre-trained model gives more than one sematic similar words in donor languages based on the given Uyghur word.

$$\{u_i, [w_{i1} : v_{i1}, w_{i2} : v_{i2}, ..., w_{ik} : v_{ik}]\} \tag{2}$$

here, k is the length of returned source word list in donor language based on cross-lingual embeddings.

Based on the distance v_i between the embeddings of a Uyghur word u_i and the first source word w_{i1} returned by the CdbCLE, we discard the Uyghur word and source words item when the distance is little than a certain threshold ρ (this value is obtained through experiments)[2]. After that, the loanword candidate list is generated.

4.3 Loanword Prediction Model

Our cross-lingual word embedding based loanword identification approach generates some candidates which are similar semantically. To predict the final outputs, we propose a log-linear model which integrate several features into our approach.

Transliteration

The string similarity feature and hybrid language model features are all rely on writing system heavily. If the receipt and donor languages belong to different writing systems, we can' t compute these two features directly. In this study, we use a transliteration model to generate a character mapping table between Uyghur and donor languages (especially for language has a different writing system with Uyghur) which make sure the receipt and donor language belong to the same writing system.

The transliteration model used in this paper can be define as

$$t_{best} = arg\max_{t} \prod_{i=1}^{I} \phi(s_i|t_i) \prod_{i=1}^{|\mathbf{t}|} p_{LM}(t_i|t_1, t_2, ..., t_{i-1}) \tag{3}$$

this model is actually a phrase-based machine translation model without reordering part. Here, we first segment the training corpus into characters and learn a phrase-based translation system over character pairs (source part:$\{s_0, s_1, s_2, ..., s_{I-1}\}$, target part:$\{t_0, t_1, t_2, ..., t_{I-1}\}$. $\phi(s_i|t_i)$ indicates the phrase translation table, $p_{LM}(t_i|t_1, t_2, ..., t_{i-1})$ means the language model probabilities (built from the target-side if mined transliteration corpus). The transliteration model assumes that source and target characters are generated monotonically. Therefore, we dont use any reordering model.

Pronunciation Similarity Feature

Usually, a loanword has the similar pronunciation to its corresponding word in the donor language. A straightforward way to identify loanword is to compute pronunciation similarity between the current word and its corresponding loanword candidates in the donor language. In this study, we use the edit distance algorithm as a feature in our prediction model.

Edit distance (ED), also known as the Levenshtein distance, is a string metric for measuring the difference between two sequences. Informally, the edit distance between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other. However, if two words belong to different word formation systems (Uyghur and Chinese), the original ED algorithm does not perform well. For example, due to the suffixes of a Uyghur word, the number of deletions according to the ED algorithm is equal or even greater than the length of donor word. Intuition might suggest that stemming of a Uyghur word can avoid such problems. However, this approach depends heavily on the performance of stemming algorithm. We propose a position-related edit distance (PRED) method , which ignores deletions at the end of word b. That is,

[2]A word with the semantic similarity lower than a certain threshold is not a loanword or is a loanword changed its meaning; we cant apply this kind of word pairs in low-resource NMT.

$$PRED_{a,b}(i,|b|) = \min\{PRED_{a,b}(i-1,|b|), ED_{a,b}(i,|b|)\} \tag{4}$$

here, a is a Uyghur word and b is a donor word, ED is the traditional edit distance algorithm.

We can define the pronunciation similarity feature function as

$$f(u,t,V) = \begin{cases} \frac{\sum_{i=1}^{k} \psi_i PRED(u,v_i)}{\sum_{i=1}^{k} len(v_i)} & if \frac{\sum_{i=1}^{k} \psi_i PRED(u,v_i)}{\sum_{i=1}^{k} len(v_i)} \leq \frac{1}{2} len(u), \\ 0 & else. \end{cases} \tag{5}$$

here, V is a set of words have similar sematic space in donor language with one candidate loanword in Uyghur u, t is the possible tag of u, $len(v_i)$ indicates the length of word v_i, ψ_i is a parameter.

Hybrid Language Model Feature

When borrowing, a word in the donor language may adopt the receipt languages pronunciation system. Therefore, the pronunciation of a loanword belongs to two different pronunciation systems (receipt and donor). Each pronunciation system can be represented by a character-level language model. This inspired us to combine two language models of pronunciation system to simulate the pronunciation of loanwords.

$$p_{hlm}(c_1,c_2,...,c_l) = (1-\lambda)p_{rec}(c_1,c_2,...,c_l) + \lambda p_{don}(c_1,c_2,...,c_l) \tag{6}$$

where $\{c_1,c_2,...,c_l\}$ is a character sequence of a candidate loanword, p_{rec} is the character level language model probability of a given character sequence in receipt language, p_{don} is the character level language model probability of a given character sequence in the donor language. λ is the prior probability of donor language model.

Accordingly, we can formulize the hybrid language model feature as:

$$f(u,t) = \begin{cases} p_{don}(u), & if p_{don}(u) \geq p_{rec}(u) \\ p_{hlm}(u), & else. \end{cases} \tag{7}$$

Loanword Prediction Model

To integrate above two features together, we use a log-linear model. One important reason is that the log-linear model allows a very rich set of features to be used in a single model, arguably much richer representations than the other machine learning algorithms.

Assume we have a candidate loanword u, and a set of possible tags T (LW/O). LW means loanword and O means not loanword. The goal of our task is to model the conditional probability $p(t|u)$. So we can describe the prediction model as:

$$p(t|u;v) = \frac{exp(v \cdot f(u,t))}{\sum_{t' \in t} exp(v \cdot f(u,t'))} \tag{8}$$

where $f(.)$ indicates a feature function, which maps (u,t) pair to a feature vector $f(u,t)$, and **v** is a parameter vector. Note that the number of features and parameters should be the same.

We use the maximum-likelihood estimation (MLE) to estimate the parameters of the log-linear model. To overcome the overfitting, we follow the common solution that modifies the objective function to include a regularization term.

4.4 Applications in OOV Translation

In this study, we integrate the results of our proposed Uyghur loanword identification model in OOV translation in NMT with following two methods:

Training data extend: we extend the training data with loanword pairs (as parallel sentences) directly.

Bilingual dictionary extend: we first extend the bilingual dictionary with our loanword pairs; then, we follow (Luong et al., 2014) to annotate the training data with explicit information that to enables the NMT system to emit, for each OOV word, a pointer to its corresponding word in the source sentence. The information is later used in post-processing to translate the OOVs words using the extended bilingual dictionary.

Type	Name	Size(Sent./Tok.)		
		Train set	Dev set	Test set
BilCorp4MT	UyAr(MT)	0.10M/2.35M	1K/23K	1.5K/30K
	UyCn(MT)	0.25M/4.79M	1K/25K	1.5K/32K
	UyRu(MT)	0.09M/2.01M	1K/19K	1.5K/28K
	UyPr(MT)	0.15M/3.08M	1K/24K	1.5K/30K
CompCorp4Train	UygMono	22.50M/450.10M(UygChn), 26.30M/478.02M(UygAra), 19.84M/405.37M(UygRus), 23.96M/500.58M(UygPer).		
	ChnMono	29.76M/502.42M	N/A	N/A
	AraMono	25.50M/469.23M	N/A	N/A
	RusMono	20.96M/402.65M	N/A	N/A
	PerMono	23.19M/486.29M	N/A	N/A

Table 1: Statistic of Corpus.

5 Experiments

To evaluate the effectiveness of our proposed approach, we conducted two groups of experiments: loanword identification and OOV translation. In loanword identification experiments, we try to detect loanwords in Uyghur in four donor languages: Chinese, Persian, Russian, and Arabic; the details can be found in Table 2. In OOV translation experiments, we trained several NMT models on eight translation directions, all of them with serious OOV problems due to the limited training data sets. We integrated the results of the loanword identification experiments into OOV translation.

5.1 Datasets

We applied two types of data in our experiments: comparable and bilingual. The details of these corpora can be found in Table 1:

BilCorp4MT:This set includes bilingual corpora for training and developing NMT translation models. We collected these corpora from Wikipedia, websites, government documents and several other resources. Also, we generated bilingual lexicons on training data in this sets. The test set of the Uyghur part was annotated manually, which include loanword information in four donor languages. Therefore, we used the test set both for evaluation of loanword identification and NMT.

CompCorp4Train:We used these data sets to train word embedding models, and apply them in cross-lingual word embedding based loanword candidate list generation approach. We extract some loanword pairs (100) to train transliteration models.

5.2 Settings

Loanword identification experiments We compared our method with other three previous proposed models for loanword identification. Before that, we finished transliteration based on Moses[3] with default settings.

MutilShlFeats:(Multiple shallow features based model): which was proposed by (Mi et al., 2014), which use pronunciation similarity, common substring to build a loanword detection model in Uyghur.

NNBased: (Neural Network based model): This method was described by (Mi et al., 2016), which relied on pronunciation similarity and a small annotated corpus to build a loanword identification model.

Ours: We applied the open source toolkit word2vec[4] to train our cross-lingual word embeddings. We set the window size as 5 and negative sampling parameter as 5. We merged the documents based on method described in section 4.2. We implemented the PRED and language model tools by ourselves.

[3]https://github.com/dav/word2vec.

[4]http://www.statmt.org/moses/.

Donor Language	Model	Results (%)		
		Precision	Recall	F1-value
Arabic	MutilShlFeats	77.45	72.59	74.94
	NNBased	79.31	73.52	76.31
	Ours	**80.82**	**74.29**	**77.42**
Chinese	MutilShlFeats	75.20	70.45	72.75
	NNBased	77.09	71.13	73.99
	Ours	**79.33**	**73.56**	**76.34**
Persian	MutilShlFeats	78.26	72.97	75.52
	NNBased	79.50	73.95	76.62
	Ours	**80.77**	**74.60**	**77.56**
Russian	MutilShlFeats	76.45	70.57	73.39
	NNBased	78.65	70.94	74.60
	Ours	**79.90**	**71.34**	**75.38**

Table 2: Evaluation on different loanword identification methods in four donor languages (ρ = 0.70).

Lang	Precision									
	ρ**=0.50**	**0.55**	**0.60**	**0.65**	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**
Arabic	0.52	0.64	0.69	0.76	**0.80**	0.58	0.53	0.49	0.41	0.36
Chinese	0.58	**0.79**	0.75	0.72	0.67	0.63	0.56	0.54	0.40	0.38
Persian	0.43	0.52	0.65	0.74	**0.80**	0.72	0.62	0.48	0.32	0.24
Russian	0.49	0.67	0.70	0.71	**0.79**	0.64	0.55	0.51	0.45	0.39

Table 3: Evaluation the affection of threshold value ρ (in **CdbCLE** model) on the precision of Uyghur loanword identification model.

The differences between our proposed model and previous studies are: 1) we used a transliteration model to obtain character mapping table for pronunciation similarity computation between two writing systems and a fixed mapping rule table was used in previous work; 2) previous loanword identification models mainly relied on large-scale annotated corpora and some rules. Our model based on a cross-lingual embedding model and loanword candidates were extracted from monolingual corpora.

OOV translation experiments

We conducted OOV translation experiments on NMT. Based on loanword identification results, we built loanword dictionaries for Uyghur-Chinese, Uyghur-Arabic, Uyghur-Russian and Uyghur-Persian. We extended bilingual corpus with these dictionaries.

NMT:We conducted NMT experiments using the open source Nematus toolkit[5]. Nematus has been used to build top-performing submissions to shared translation tasks at WMT and IWSLT. In this study, we use the default settings: the word embedding dimension was 620 and the size of a hidden layer was 1000, the vocabulary size was 30K, the batch size was 80, the maximum sequence length was 50, and the beam size for decoding was 10. Default dropout was applied. Each NMT model was trained for 80 batches by using AdaDelta optimizer (Zeiler, 2012).

5.3 Results and Analysis

Corpora	Size (word pair)			
	Uyghur-Arabic	Uyghur-Chinese	Uyghur-Persian	Uyghur-Russian
Dictionary	62,820	77,592	54,902	80,367
Dictionary (extend)	**63,652** (+832)	**78,342** (+750)	**55,408**(+506)	**81,065**(+698)

Table 4: Size of bilingual dictionaries (before and after extend).

[5]https://github.com/EdinburghNLP/nematus.

Language pair	BLEU			OOV rate		
	sys1	**sys2**	**sys3**	**sys1**	**sys2**	**sys3**
Uyghur-Russian	15.94	16.71	**16.93**	18.25	17.08	**16.94**
Uyghur-Persian	17.88	18.25	**18.47**	18.63	17.27	**17.10**
Uyghur-Chinese	22.50	22.94	**23.15**	18.05	17.48	**17.29**
Uyghur-Arabic	19.79	20.03	**20.31**	18.69	18.25	**17.84**
Russian-Uyghur	13.51	13.84	**14.07**	19.07	18.69	**18.52**
Persian-Uyghur	16.14	16.59	**16.80**	18.92	18.73	**18.56**
Arabic-Uyghur	17.82	18.23	**18.52**	19.97	19.45	**19.29**
Chinese-Uyghur	19.26	19.85	**20.03**	19.32	18.74	**18.61**

Table 5: Experiment results on OOV translation of different language pairs. **sys1** indicates the BPE based model (baseline), **sys2** means BPE based model with loanword pairs in training data, **sys3** means BPE based model with loanword pairs in post-processing step (Section 4.4).

In Table 2, we present loanword identification results in Uyghur. We find that across the four donor languages, our proposed method achieves the best performance compares with other three models. One important reason is that our method combines both semantic similarity (cross lingual word embedding) and pronunciation similarity (edit distance and language model). Among different donor languages, experimental results show that Persian loanwords in Uyghur can be identified more effectively compared with Arabic, Chinese and Russian. This is because Uyghur and Persian share much more words and has the most similar word formation. The loanword identification of Arabic also performs well compare with Russian and Arabic, the most possible reason is that many loanwords entered the language directly through Islamic literature. Although many loanwords of Europe origin have reached Uyghur through Russia's influence, the results of Russian loanword identification cannot outperform others. We believe that relatively fewer corpora lead to this situation. Due to the significant different word formation, the Chinese loanword detection achieves the worst results.

We find that the precision of our proposed model rely on threshold value heavily (Table 3). There are two possible reasons lead to this situation: size of data and diversity of languages. Since we have different size of data for each language pairs and the performance of cross-lingual embedding model usually based on the data, our model achieves best precision of four donor languages in different thresholds (0.55 and 0.70). Another reason is diversity of languages. We can easily find that Arabic, Persian and Russian are all achieves best results when ρ is **0.70**, only Chinese achieves its best precision score when ρ is **0.55**. After analysis, we know that Russian, Arabic and Persian have the similar word formation system with Uyghur, which is very different from Chinese.

We evaluate translation quality both with the BLEU score and OOV rate in NMT experiments (Table 5). We find that in all NMT tasks, models optimized by our loanword identification experiments received significant improvements compared with baseline methods. After extending the bilingual corpus by our loanword identification results, the OOV rates decrease significantly. Since BPE has achieved many state-of-the-art results in recent studies, we choose the BPE based model as the baseline (**sys1**). Due to the extend size of data sets, systems with loanword pairs are all outperform the baseline model, but with different BLEU improvements. For **sys2** (BPE+loanwordpairs+train), we use the loanword pairs in model training as other parallel sentences. Therefore, some OOV may not translate due to data spareness. **sys3** overcome this problem through translate OOV words in post-processing step with extended dictionaries Table 4. So the sys3 achieves best performance among all three translation systems in all language pairs.

6 Conclusion

In this study, we try to identify loanwords in low-resource languages (Uyghur) to extend bilingual corpus. In our proposed method, we first use the cross-lingual word embeddings as the core feature to generate semantically related candidates based on large comparable corpora and a small bilingual lexicon; then,

a log-linear model which combines several shallow features is explored to predict the final results. Results of loanword identification suggest that our proposed model outperforms other baseline methods significantly. Also, we conduct experiments on OOV translation in NMT. Experimental results show that loanword pairs can help reduce OOV rates in several low-resource language pairs.

In future work, we plan to import transfer learning into our model to further improve the performance.

Acknowledgements

We are grateful to the mentor of this paper for his meaningful feedback. Thanks Rui Wang and Kehai Chen for helpful discussions and three anonymous reviewers for their insightful comments and suggestions. This work is supported by the West Light Foundation of The Chinese Academy of Sciences under Grant No.2015-XBQN-B-10, Xinjiang Science and Technology Major Project under Grant No.2016A03007-3 and National Natural Science Foundation of China (NSFC) under Grant No.U1703133.

References

Oliver Adams,Adam Makarucha,Graham Neubig,Steven Bird and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics,* pages 937–947, Valencia, Spain.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, Noah A. Smith. 2016. Massively multilingual word embeddings. *arXiv 2016.*

Long Duong,Hiroshi Kanayama,Tengfei Ma,Steven Bird and Trevor Cohn. 2017. Multilingual Training of Crosslingual Word Embeddings. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics,* pages 894–904, Valencia, Spain.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics,* pages 587–593, Vancouver, Canada.

Stephan Gouws, Yoshua Bengio and Greg Corrado. 2015. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. *Proceedings of the 32nd International Conference on Machine Learning,* pages 748–756, Lille, France.

Bates L. Hoffer. 2005. Language borrowing and the indices of adaptability and receptivity. *Intercultural communication studies*, 14(2):53–72.

Yoonjung Kang, Andrea H Phạm, and Benjamin Storme. 2014. French loanwords in Vietnamese: the role of input language phonotactics and contrast in loanword adaptation. *Proceedings of the Annual Meetings on Phonology,* MIT.

Alexandre Klementiev, Ivan Titov and Binod Bhattarai. 2012. Inducing Crosslingual Distributed Representations of Words. *Proceedings of COLING 2012,* pages 1459–1474, Mumbai, India.

Nicholas Kontovas. 2008. An analysis of recent loans into the Standard Uyghur lexicon. *University of Chicago,* Chicago.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals and Wojciech Zaremba. 2014. Addressing the Rare Word Problem in Neural Machine Translation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing,* pages 11–19, Beijing, China.

Chenggang Mi, Yating Yang, Lei Wang, Xiao Li and Kamali Dalielihan. 2014. Detection of Loan Words in Uyghur Texts. *Proceedings of the 3rd International Conference on Natural Language Processing and Chinese Computing,* pages 103–112, Shenzhen, China.

Chenggang Mi, Yating Yang, Xi Zhou, Lei Wang, Xiao Li and Tonghai Jiang. 2016. Recurrent Neural Network Based Loanwords Identification in Uyghur. *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation,* pages 209–217, Seoul, Korea.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems 26,* pages 3111–3119, Lake Tahoe, California.

Toan Q. Nguyen and David Chiang. 2017. Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation. *Proceedings of the Eighth International Joint Conference on Natural Language Processing,* pages 296–301, Taipei, Taiwan.

Sharon Peperkamp. 2004. A psycholinguistic theory of loanword adaptations. *Proceedings of the 30th Annual Meeting of the Berkeley Linguistics Society,* pages 341–352, Berkeley, CA.

Sebastian Ruder, Ivan Vulić and Anders Søgaard. 2017. A Survey Of Cross-lingual Word Embedding Models. *arXiv 2017.*

Rico Sennrich, Barry Haddow and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* pages 1715–1725, Berlin, Germany.

Shigeko Shinohara. 2015. Loanword-specific grammar in Japanese adaptations of Korean words and phrases. *Journal of East Asian Linguistics*, 24(2): 149-191.

Alexander Sugar. 2017. Mandarin Chinese Verbs as Verbal Items in Uyghur Mixed Verbs. *Languages*, 2(1), 1.

Yulia Tsvetkov, Waleed Ammar and Chris Dyer. 2015. Constraint-Based Models of Lexical Borrowing. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* pages 598–608, Denver, Colorado.

Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui,Guillaume Lample,Patrick Littell,David Mortensen,Alan W Black,Lori Levin and Chris Dyer. 2016. Polyglot Neural Language Models: A Case Study in Cross-Lingual Phonetic Representation Learning. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* pages 1357–1366, San Diego, California.

Yulia Tsvetkov and Chris Dyer. 2016. Cross-lingual bridges with models of lexical borrowing . *Journal of Artificial Intelligence Research*, 55(1): 63-93.

Yulia Tsvetkov and Chris Dyer. 2016. Lexicon Stratification for Translating Out-of-Vocabulary Words. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing,* pages 125–131, Beijing, China.

Ivan Vulić and Marie F Moens. 2015. Bilingual Word Embeddings from Non-Parallel Document-Aligned Data. Applied to Bilingual Lexicon Induction.. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing,* pages 719–725, Beijing, China.

Liangchen Wei and Zhihong Deng. 2017. A Variational Autoencoding Approach for Inducing Cross-lingual Word Embeddings. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence,* pages 4165–4171, Melbourne, Australia.

Matthew D. Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *arXiv 2012.*