A Neural Network Based Model for Loanword Identification in Uyghur

Chenggang Mi[†‡], Yating Yang[†‡], Lei Wang[†‡], Xi Zhou[†‡], Tonghai Jiang[†‡]
[†]The Xinjiang Technical Institute of Physics & Chemistry of Chinese Academy of Sciences, Urumqi, China
[‡]Key laboratory of speech language information processing of Xinjiang, Urumqi, China
{micg,yangyt,wanglei,zhouxi,jth}@ms.xjb.ac.cn

Abstract

Lexical borrowing happens in almost all languages. To obtain more bilingual knowledge from monolingual corpora, we propose a neural network based loanword identification model for Uyghur. We build our model on a bidirectional LSTM - CNN framework, which can capture past and future information effectively and learn both word level and character level features from training data automatically. To overcome data sparsity that exists in model training, we also suggest three additional features , such as hybrid language model feature, pronunciation similarity feature and part-of-speech tagging feature to further improve the performance of our proposed approach. We conduct experiments on Chinese, Arabic and Russian loanword detection in Uyghur. Experimental results show that our proposed method outperforms several baseline models.

Keywords: Loanword Identification, Uyghur Language, BiLSTM-CNN, Language Model, Pronunciation Similarity

1. Introduction

Lexical borrowing is very common between languages (Taylor and Grant, 2015). It is a phenomenon of cross-linguistic influence. If loanwords in resource-poor languages (e.g. Uyghur) can be identified effectively, we can use the donor-receipt word pairs to extend bilingual dictionary. And the bilingual dictionary plays a very important role in many cross-lingual areas in natural language processing (NLP), such as machine translation (Tsvetkov and Dyer, 2015).

In this paper, we describe a novel method to identify loanwords in Uyghur texts to alleviate the data sparsity that exists in Uyghur related NLP tasks. Our loanword identification model is based on a bidirectional long-short term memory (BiLSTM) - convolutional neural networks (CNNs) framework (Chiu and Nichols, 2016). The BiLSTM have achieved state-of-the-art performance in varies sequence-to-sequence learning tasks, a very important reason is that it can capture past (from previous tagged words) and future (from next untagged words) information effectively. We use it to model word level features. CNNs have been used in several character level natural language processing (NLP) tasks; we use it to model character level features. Therefore, our model can learn both word level and character level feature from training data, automatically. Additionally, we also propose three features (hybrid language model, pronunciation similarity and part-of-speech tagging) to argument the BiLSTM-CNN model by exploit knowledge learned from monolingual corpus. We conduct experiments on Chinese, Arabic and Russian loanwords detection in Uyghur, respectively. Experimental results show that our model outperforms several baseline models.

2. Loanwords in Uyghur Language

2.1. An Introduction of Uyghur Language

Uyghur is an official language of the Xinjiang Uyghur Autonomous Region in China, and is widely used in both social and official spheres, as well as in print, radio and television, and is mostly used as a lingua franca by other ethnic minorities in Xinjiang. Uyghur belongs to the Turkic language family. Like other Turkic languages, Uyghur displays vowel harmony and agglutination, lacks noun classes or grammatical gender, and is a left-branching language with subject (S) - object (O) - verb (V) word order.

As an agglutinative language, nouns in Uyghur are inflected for number and case, but not gender and definiteness like in many other languages(Table 1)[1]. There are two numbers (singular, plural) and six different cases in nouns of Uyghur language. Verbs are inflected for tense, voice, aspect and mood.

Uyghur(in English)	Uyghur(stem+suffix)
aliqanimda(In my hands)	**aliqan**+im+da
etrapidikilerni(People around)	**etrap**+i+diki+ler+ni
qurulmasining(Structured)	**qurulma**+si+ning
qalduridu(Stay)	**qal**+dur+i+d+u

Table 1: Examples of Uyghur word formation.

2.2. Linguistic Issues of Loanwords in Uyghur

Due to different kinds of language contact through the history, Uyghur has adopted many loanwords(Kamalov, 2006). Some studies show that larger than 20% of the vocabulary is from other languages. Kazakh, Uzbek, and Chagatai are all Turkic languages which had a strong influence on Uyghur. Arabic words have also entered Uyghur through Islamic literature after the introduction of the Islamic religion around the 10th century.

Recently, Chinese and Russian had the greatest influence on Uyghur language. In particular, loanwords from these two languages are all quite recent. Below are some examples of loanwords in Uyghur (Table 2):

3. Method

In this section, we describe the details of our proposed loanword identification model. First, we present the BiL-

[1]In this paper, we write Uyghur with the Latin alphabet.

Uyghur(Chinese)[In English]	Uyghur(Russian)[In English]
shinjang(新疆) [one province in China]	tElEfon(телефон) [telephone]
laza(辣子) [hot pepper]	uniwErsitEt(университет) [university]
shuji(书记) [secretary]	radiyo(радио) [radio]
koi(块) [Chinese currency]	pochta(почта) [post office]
lengpung(凉粉) [agar-agar jelly]	wElsipit(велосипед) [bicycle]
dufu(豆腐) [tofu]	oblast(область) [region]

Table 2: Examples of Chinese and Russian Loanwords in Uyghur.

STM model, which can model word-level features from both forward and backward directions. Next, we introduce the CNN based character-level feature extraction. Then, we detail two core features (word embedding feature and character embedding feature) and three additional features (pronunciation similarity feature, POS tagging feature and hybrid language model feature) used in our model. Finally, we present the training and optimization of our proposed neural network.

3.1. Word-level Tagging with BiLSTM

Similar to previous work used the BiLSTM in other areas in NLP (Such as NER and speech recognition), we explored a stacked bi-directional recurrent neural network with LSTM units to transform word features into loanword tag scores(Figure 1).

In our paper, we fed extracted features of each word into

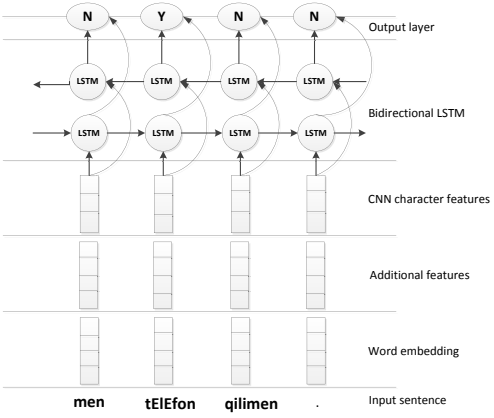

Figure 1: BiLSTM Model for Loanwords Identification in Uyghur.

a forward and backward LSTM network. The output of each network at each time step is decoded by a linear layer and a log-softmax layer into log-probabilities for each tag category. The two vectors produced by a linear and a log-softmax are added to produce the final output.

3.2. Character-level Features Extraction Based on CNN

Convolutional neural networks (CNN) have shown great success in character level features extraction in NER and POS tagging. However, the character level CNN has not been applied in loanwords identification. In this paper, we employ a convolution and a max layer to extract a new feature vector from the character embeddings(Figure 2).

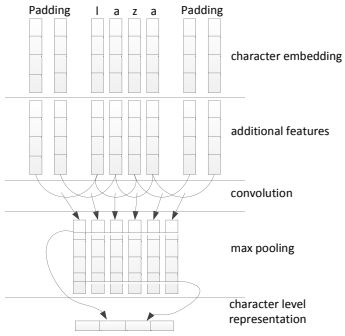

Figure 2: Character-level Features Extraction Model Based on CNN.

3.3. Features Definition

3.3.1. Main Features

Word Embedding Feature

Word embedding is a collective name for techniques in NLP where words or phrases from the vocabulary are mapped to vectors represent by real numbers. It aims to quantify and categorize semantic similarities between words based on their distributional properties in large samples of language data. In this study, we use the word embedding as input to the neural network.

Currently, there is no publicly available Uyghur word embeddings, so we trained them by ourselves. In this paper, we experimented with three sets of word embeddings: 1) NEWS word2vec embeddings trained on 2 billion words from news and government documents; 2) ORAL word2vec embeddings trained on 1.5 billion words from short messages and weinxin, 3) HYBRID word2vec embeddings trained on 3.5 billion words from both 1) and 2). We used the open source toolkit Glove[2] to train above word embeddings.

Character Embedding Feature

In this paper, the character embeddings are uniformly sampled from range $[-\sqrt{\frac{3}{DIM}}, \frac{3}{DIM}]$. DIM is the dimension of embeddings. The character set includes all unique characters in Uyghur language. Besides, there are two more tokens are also containing in above set: UNKNOWN and PADDING. The UNKNOWN is indicates all other characters and PADDING is used for CNN.

[2]https://nlp.stanford.edu/projects/glove/

3.3.2. Additional Features

Besides the main features described in previous section, we also proposed four features to further improve the performance of our approach. Intuitively, we may think that the loanword in Uyghur should have a similar pronunciation with its corresponding word in donor language. Therefore, we take the pronunciation similarity as an important feature in our model. Due to most loanwords are nouns, we constraint the output of our approach by a part-of-speech tag feature. A loanword may adapt the pronunciation system of receipt language when borrowing from the donor language. So the pronunciation of a loanword has the features of pronunciation systems of both receipt and donor languages. We use a hybrid language model to represent this feature. In this paper, we focus on Chinese, Arabic and Russian loanwords in Uyghur.

Pronunciation Similarity (ps)

Loanword (LW) usually have a similar pronunciation to their corresponding donor language (DW) word. Previous work detected loanwords in Uyghur according to string similarity between LW and DW, which first transfer the pronunciation similarity as string similarity. However, we found that there exist many differences among different writing systems, which have negative effects on the loanwords detection. To overcome this, we proposed a method that transfers both words into strings according to the International Phonetic Alphabet (IPA) . After that, we compute string similarity based on the Edit Distance algorithm (aka Levenshtein Distance).

Part-Of-Speech Tagging (pos)

Most loanwords are nouns. Therefore, we use part-of-speech information as one of the additional feature to constraint the output of our model. The POS tags are obtained by an in-house Uyghur POS tagger which was developed by us.

Hybrid Language Model (hlm)

Usually, there are different pronunciation systems between recipient language and donor language. Each pronunciation system can be represented by a certain character-level language model. When lexical borrowing, the pronunciation of a word in donor language may adapts the pronunciation system of the recipient language. This inspired us to combine these two language models to feature the pronunciation of loanwords.

$$p_{hlm} = (1 - \lambda_1 p_{uyg}) + \lambda_2 p_{dnr} \quad (1)$$

Where p_{uyg} is the language model probability of a given character sequence in Uyghur, p_{dnr} is the language model probability of above sequence in donor languages. λ_1 and λ_2 are weights which can be obtained during model optimization.

3.4. Neural Network Training

3.4.1. Tagging Scheme

Similar to the scheme used in named entities recognition (NER), we used the BIESO tag set in this paper. BIOES, which stand for Begin (B), Inside (I), End (E), Single (S) and Other (O), indicating the position of the character in a certain loanword. For example: With the BIESO tagging scheme, more information can be considered when neural networks training.

3.4.2. Network Implementation

We implement the BiLSTM-CNN model used in our paper based on Theano[3], a widely used deep learning Python library. We trained the loanwords identification model based on sentence-level corpus. We initialized the word embedding and character embedding as previous description. Other lookup tables used in our model are randomly initialized with values drawn from a standard normal distribution.

3.4.3. Parameters

We tune the hyper-parameters as follows: for CNN, we set the window size as 5 and use 40 filters; for bidirectional LSTM network, we set the initial state as 0.0 and state size as 300. As mentioned above, we use dropout to regularize our model to alleviate overfitting when neural network training and we set the dropout rate as 0.5.

3.4.4. Optimization Algorithm

We use the min-batch stochastic gradient descent (SGD) to train our loanwords identification model. Each minibath includes multiple sentences with the same number of Uyghur words. We find that training neural network with dropout is very effective in alleviate the overfitting. To achieve a better performance on development sets, we use early stopping method in our experiments.

4. Experiments

4.1. Datasets and Settings

To evaluate the proposed method effectively, we train the neural network with three groups of corpora (tokens)(UYGLWChn:20M/10K/20K*2 [training/develop/test], UYGLWArab:15M/10K/20K*2 and UYGLWRus:12M/10K/20K*2). Because there are relatively few loanwords (compared with other words) in Uyghur, we also annotated person names in these three donor languages as loanwords. These corpora are collected from government websites and newspapers. Test data for cross-domain experiments are selected from social medias such as Weixin and Twitter. We train three donor language models with corpus selected from previous corpora. The develop sets and test sets used in our experiments are all selected from the same domain.

We built the bi-directional LSTM-CNN neural network on the Theano library. The computations for a model are run on a GPU. We extract the word embedding and character embedding based on the open source toolkit Glove[4]. We use SRILM[5] to obtain the character level language model for four languages. The POS tagging features are extracted based on a Uyghur POS tagger, which was developed by us.We use Precision (P), Recall (R) and F1 score to evaluate the performance of loanword identification models.

[3] http://deeplearning.net/software/theano/

[4] https://nlp.stanford.edu/projects/glove/

[5] http://www.speech.sri.com/projects/srilm/

Model	Pchn	Rchn	F1chn	Prus	Rrus	F1rus	Parab	Rarab	F1arab
CRFs	69.78	62.33	66.35	71.64	63.25	67.18	72.50	65.32	68.72
SSIM	66.32	77.28	71.38	75.39	70.02	72.61	73.76	67.51	70.50
CIBM	78.82	68.30	73.18	81.03	73.22	76.93	75.22	70.71	72.90
RNN	78.97	79.20	79.08	82.55	75.93	79.10	83.26	77.58	80.32
Ours	**80.24**	**81.02**	**80.63**	**82.95**	**76.30**	**79.49**	**84.09**	**78.28**	**81.08**

Table 3: Experimental Results on Loanword Identification Models.

4.2. Experimental Results

4.2.1. Effects on Different Methods

Table 3 presents results on different methods, including CRFs-based model (CRFs), the string similarity based loanword identification model (SSIM) (Mi et al., 2013), identification model based on classification (CBIM) (Mi et al., 2014), a RNN based identification model (Mi et al., 2016) and the model proposed in this paper. Due to lack sufficient annotated corpus, CRFs based model cannot outperform other models. After analysis the output of CRFs based model, we found that only a small number of non-person name loanwords are identified, which means the CRFs based method rely on annotated corpus heavily. The performance of SSIM outperforms CRFs based model, an important reason is that the SSIM method based on pronunciation similarity between two words (donor words and receipt word). Because it combines the advantages of statistical based and rule based model, the CIBM outperforms CRFs and SSIM models. Like CRFs, RNN based model also rely on annotated corpus heavily. However, we found that the RNN based model can identify more loanwords (more than person names) than CRFs based model, a possible reason is that the RNN encoder-decoder framework can learn features automatically and use its internal memory to process arbitrary sequences of inputs. To model the word-level and character-level features, we proposed a BiLSTM-CNN neural network to identify loanwords from Uyghur texts. Moreover, three important additional features were also suggested to overcome data sparseness. Our proposed method achieved the best score among these approaches.

4.2.2. Experimental Results on Different Features

In Table 4, we present results with different additional features. We can found that our model (ours) achieves best performance among all these models. We observe that our model benefit most from hlm (**h**ybrid **l**anguage **m**odel) feature. Compare with other two models (BiLSTM-CNN+ps&pos), model with hlm feature achieve the best improvements. An important reason is that the hlm feature combines both pronunciation systems of donor language and receipt language. Models with string similarity and POS tag features only reflect the shallow semantic information. Due to lack of annotated corpus, the model without any additional features performs worst in all experiments.

4.2.3. Evaluation on Different Domains

Table 5 shows our results on different domains. To show the capability of our loanwords identification model, we evaluate our model on different domains, such as news(News) and social network(socialNet). We can found that the experimental results on formal corpus (news domain) which have the same domain with our training corpus are outperforming the performance on informal domain (social network). We observed the same situation in all donor languages. We also found that experimental results on informal domain are just a little worse compared with results on formal domain. One possible reason is that our BiLSTM+CNN model can learn representation of knowledge beyond given training examples.

5. Related work

In general, word borrowing is often concerned by linguists (Chen, 2011)(Chen and Chen, 2011). There are relatively few studies about loanwords in NLP area.(Tsvetkov et al., 2015) and (Tsvetkov and Dyer, 2016) proposed a morphphonological transformation model, features used in this model are based on optimality theory; experiment has been proved that with a few training examples, this model can obtain good performance at predicting donor forms from borrowed forms.(Tsvetkov and Dyer, 2015) suggest an approach that uses the lexical borrowing as a model in SMT framework to translate OOV words in a low-resource language. For loanwords detection in Uyghur, string similarity based methods were often used at the early stage(Mi et al., 2014).(Mi et al., 2016) propose a loanword detection method based on the perceptron model, several features are used in model training.

6. Conclusion

In this paper, we have presented a novel model to detect loanwords (mainly Chinese, Arabic and Russian loanwords) in Uyghur by using a BiLSTM-CNN framework. Except two main features such word embeddings and character embeddings, two additional features like donor language model feature and hybrid language model feature are also proposed and integrated the framework to further improve the performance. In the proposed model, the character-level feature of each word is extracted by the CNN model based on character embedding and our proposed two features. For each word, the character-level feature vector is concatenated with the word embedding feature vector and fed into the BiLSTM model. After that, these feature vectors are fed to output layers. Experiment results on loanwords identification in Uyghur have presented that the proposed model can significantly improve the identification performance.

Although our model achieves the best results on loanwords identification in Uyghur, we only use loanword taggers in our training set. In our future work, we plan to integrate

Feature	Pchn	Rchn	F1chn	Prus	Rrus	F1rus	Parab	Rarab	F1arab
BiLSTM-CNN	77.65	67.89	72.44	78.02	68.33	72.85	78.38	70.96	74.49
BiLSTM-CNN+ps	78.86	70.32	74.35	81.94	70.65	75.88	81.12	71.52	76.02
BiLSTM-CNN+pos	78.79	69.54	73.88	81.35	71.28	75.98	80.76	70.20	75.11
BiLSTM-CNN+hlm	79.42	70.37	74.62	82.29	73.50	77.65	82.14	73.59	77.63
BiLSTM-CNN+all	**80.24**	**81.02**	**80.63**	**82.95**	**76.30**	**79.49**	**84.09**	**78.28**	**81.08**

Table 4: Evaluation on Features Used in RNN-based Model.

Domain	Pchn	Rchn	F1chn	Prus	Rrus	F1rus	Parab	Rarab	F1arab
socialNet	79.63	80.51	80.07	81.05	75.22	78.03	83.63	77.45	80.42
News	**80.24**	**81.02**	**80.63**	**82.95**	**76.30**	**79.49**	**84.09**	**78.28**	**81.08**

Table 5: Evaluation on Cross-Domain Corpora.

more linguistic knowledge to further optimize the performance of our proposed model.

7. Acknowledgements

We sincerely thank the anonymous reviewers for their thorough reviewing and valuable suggestions. This work is supported by the West Light Foundation of The Chinese Academy of Sciences under Grant No.2015-XBQN-B-10, the Xinjiang Key Laboratory Fund under Grant No.2015KL031, the Xinjiang Science and Technology Major Project under Grant No.2016A03007-3 and the Natural Science Foundation of Xinjiang under Grant No.2015211B034.c

8. Bibliographical References

Chen, Y. and Chen, P. (2011). A comparison on the methods of uyghur and chinese loan words. *Journal of Kashgar Teachers College*, 32(2):51–55.

Chen, S. (2011). New research on chinese loanwords in the uyghur language. *N.W.Journal of Ethnology*, 28(1):176–180.

Chiu, J. and Nichols, E. (2016). Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Kamalov, A. (2006). *Uyghur Studies in Central Asia: A Historical Review*.

Mi, C., Yang, Y., Zhou, X., Li, X., and Yang, M. (2013). Recognition of chinese loan words in uyghur based on string similarity. *Journal of Chinese Information Processing*, 27(5):173–179.

Mi, C., Yang, Y., Wang, L., Li, X., and Dalielihan, K. (2014). Detection of loan words in uyghur texts. In *Natural Language Processing and Chinese Computing: Third CCF Conference, NLPCC 2014, Shenzhen, China, December 5-9, 2014. Proceedings*, pages 103–112, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mi, C., Yang, Y., Zhou, X., Wang, L., Li, X., and Jiang, T. (2016). Recurrent neural network based loanwords identification in uyghur. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*, pages 209–217.

Taylor, J. R. and Grant, A. P. (2015). Lexical borrowing.

Tsvetkov, Y. and Dyer, C. (2015). Lexicon stratification for translating out-of-vocabulary words. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 125–131. Association for Computational Linguistics.

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

Tsvetkov, Y., Ammar, W., and Dyer, C. (2015). Constraint-based models of lexical borrowing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 598–608. Association for Computational Linguistics.