Recurrent Neural Network Based Loanwords Identification in Uyghur

Chenggang Mi[1,2], **Yating Yang**[1,2,3], **Xi Zhou**[1,2], **Lei Wang**[1,2], **Xiao Li**[1,2] **and Tonghai Jiang**[1,2]
[1]Xinjiang Technical Institute of Physics and Chemistry of Chinese Academy of Sciences, Urumqi, 830011, China
[2]Xinjiang Laboratory of Minority Speech and Language Information Processing, Xinjiang Technical Institute of Physics and Chemistry of Chinese Academy of Sciences, Urumqi, 830011, China
[3]Institute of Acoustics of the Chinese Academy of Sciences, Beijing, 100190, China
{micg,yangyt,zhouxi,wanglei,xiaoli,jth}@ms.xjb.ac.cn

Abstract

Comparable corpus is the most important resource in several NLP tasks. However, it is very expensive to collect manually. Lexical borrowing happened in almost all languages. We can use the loanwords to detect useful bilingual knowledge and expand the size of donor-recipient / recipient-donor comparable corpora. In this paper, we propose a recurrent neural network (RNN) based framework to identify loanwords in Uyghur. Additionally, we suggest two features: inverse language model feature and collocation feature to improve the performance of our model. Experimental results show that our approach outperforms several sequence labeling baselines.

1 Introduction

Most natural language processing (NLP) tools rely on large scale language resources, but many languages in the world are resource-poor. To make these NLP tools widely used, some researchers have focused on techniques that obtain resources of resource-poor languages from resource-rich languages using parallel data for NLP applications such as syntactic parsing, word sense tagging, machine translation, semantic role labeling, and some cross-lingual NLP tasks. However, high quality parallel corpora are expensive and difficult to obtain, especially for resource-poor languages like Uyghur.

Lexical borrowing is very common between languages. It is a phenomenon of cross-linguistic influence (Tsvetkov et al., 2015a). If loanwords in resource-poor languages (e.g. Uyghur) can be identified effectively, we can use the bilingual word pairs as an important factor in comparable corpora building. And comparable corpora are vital resources in parallel corpus detection (Munteanu et al., 2006). Additionally, loanwords can be integrated into bilingual dictionaries directly. Therefore, loanwords are valuable to study in several NLP tasks such as machine translation, information extraction and information retrieval.

In this paper, we design a novel model to identify loanwords (Chinese, Russian and Arabic) from Uyghur texts. Our model based on a RNN Encoder-Decoder framework (Cho et al., 2014). The Encoder processes a variable length input (Uyghur sentence) and builds a fixed-length vector representation. Based on the encoded representation, the decoder generates a variable-length sequence (Labeled sequence). To optimize the output of decoder, we also propose two important features: inverse language model feature and collocation feature. We conduct three groups of experiments; experimental results show that, our model outperforms other approaches.

This paper makes the following contributions to this area:

- We introduce a novel approach to loanwords identification in Uyghur. This approach increases F1 score by 12% relative to traditional approach on the task of loanwords detection.
- We conduct experiments to evaluate the performance of off-the-shelf loanwords detection tools trained on news corpus when applied to loanwords detection. By utilizing in-domain and out-of-domain data.

- For integrate these crucial information for better loanwords prediction, we combine two features into the loanwords identification model, so that we can use more important information to select the better loanword candidate.

The rest of this paper is organized as follows: Section 2 presents the background of loanwords in Uyghur; Section 3 interprets the framework used in our model; Section 4 introduces our method in detail. Section 5 describes the experimental setup and the analysis of experimental results. Section 6 discusses the related work. Conclusion and future work are presented in Section 7.

2 Background

Before we present our loanwords detection model, we provide a brief introduction of Uyghur and loanwords identification in this section. This will help build relevant background knowledge.

2.1 Introduction of Loanwords

A loanword is a word adopted from one language (the donor language) and incorporated into a different, recipient language without translation. It can be distinguished from a calque, or loan translation, where a meaning or idiom from another language is translated into existing words or roots of the host language. When borrowing, the words may have several changes to adopt the recipient language:

- Changes in meaning. Words are occasionally improved with a different meaning than that in the donor language
- Changes in spelling. Words taken into different recipient languages are something spelled as in the donor language. Sometimes borrowed words retain original (or near-original) pronunciation, but undergo a spelling change to represent the orthography of the recipient language.
- Changes in pronunciation. In cases where a new loanword has a very unusual sound, the pronunciation of the word is radically changed.

2.2 Loanwords in Uyghur

Uyghur is an official language of the Xinjiang Uyghur Autonomous Region, 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, which also includes languages such as the more distantly related Uzbek. In addition to influence of other Turkic languages, Uyghur has historically been influenced strongly by Person and Arabic and more recently by Mandarin Chinese and Russian (Table 1).

Loanwords in Uyghur not only include named entities such as person and location names, but also some daily used words.

2.3 Challenges in Loanwords Identification in Uyghur

Spelling Change When Borrowed From Donor Languages

To adopt the pronunciation and grammar in Uyghur, spelling of words (loanwords) may change when borrowed from donor languages. Changes of spelling have a great impact on the loanwords identification task.
Russian loanwords in Uyghur:
"radiyo"[1]-"радио"("radio")
Chinese loanwords in Uyghur:
"koi"-"块" ("kuai")

Suffixes of Uyghur Words Affect the Loanwords Identification

A Uygur word is composed of a stem and several suffixes, which can be formally described as:

$$Word = stem + suffix_0 + suffix_1 + ... + suffix_N \quad (1)$$

If we just use the traditional approaches such as edit distance, in some cases, these algorithms cannot give us sure results, for example, the length of suffixes equal even greater than the original word's length.

Data Sparsity Degrades the Performance of Loanwords Identification Model

[1] In this paper, we use Uyghur Latin Alphabet.

Chinese loan words in Uyghur [in English]	Russian loan words in Uyghur [in English]
shinjang(新疆) [Xinjiang]	tElEfon(телефон) [telephone]
laza(辣子) [hot pepper]	uniwErsitEt(университет) [university]
shuji(书记) [secretary]	radiyo(радио) [radio]
koi(块) [Yuan]	pochta(почта) [post office]
lengpung(凉粉) [agar-agar jelly]	wElsipit(велосипед) [bicycle]
dufu(豆腐) [bean curd]	oblast(область) [region]

Table 1: Examples of Chinese and Russian Loanwords in Uyghur.

Loanwords detection can be reformulated as a sequence labeling problem. Most sequence labeling tools (such as CRFs-based, HMM-based etc.) are built on large scales labeled data, lack of available labeled language resource makes decrease of performance on loanwords identification in Uyghur using above "off-the-shelf" tools.

3 Methodology

Recent development of deep learning (representation learning) has a strong impact in the area of NLP (natural language processing). According to traditional approaches, extraction of features often requires expensive human labor and often relies on expert knowledge, and these features usually cannot be expended in other situations. The most exciting thing of deep learning is that features used in most traditional machine learning models can be learned automatically.

In this section, we first introduced the most popular deep learning models used in this paper, then, we involved in the details of this model.

3.1 Recurrent Neural Network

RNNs (Recurrent Neural Networks) are artificial neural network models where connections between units form a directed cycle (Jaeger, 2002). This creates an internal state of the network which allows it to exhibit dynamic temporal behavior. RNNs can use their internal memory to process arbitrary sequences of inputs. This makes them show great promise in many NLP tasks.

The most important feature of a RNN model is that the network contains at least one feed-back connection, so the activations can flow round in a loop. That makes the networks very suited for tasks like temporal processing and sequence labeling.

3.2 RNN Encoder-Decoder Framework

In this section, we give a brief introduction of the RNN Encoder-Decoder framework, which was proposed by (Cho et al., 2014a) and (Sutskever et al., 2014). We build a novel architecture that learns to identify loanwords in Uyghur texts based on this framework.

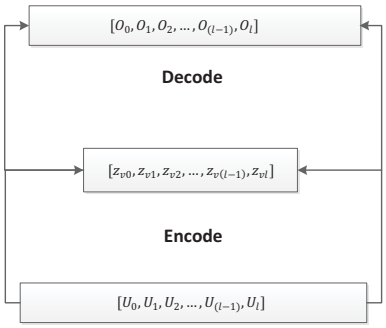

Figure 1: The Encoder-Decoder Framework Used in Loanword Identification Model.

$U_0, U_1, U_2, ..., U_{l-1}, U_l$ is a sequence of Uyghur words, $O_0, O_1, O_2, ..., O_{l-1}, O_l$ is a sequence of labels (loanword or not), and $Z_{v0}, Z_{v1}, Z_{v2}, ..., Z_{v(l-1)}, Z_{vl}$ is a sequence of vector representation of Uyghur words. The bold face "Encode" and "Decode" are two processes of encoder and decoder in our loanword identification model, respectively (Figure 1).

Encoder

In the RNN Encoder-Decoder framework, a sentence is firstly transformed into a sequence of vectors $x = (x_1, x_2, ..., x_l)$, then the encoder reads x as a vector $\vec{c}$. The most common approach is to

use an RNN such that

$$h_t = f(x_t, h_{t-1}) \tag{2}$$

And

$$c = q(h_1, h_2, ..., h_{l-1}, h_l) \tag{3}$$

where $h_t \in \gamma$ is a hidden state at time t, and $\vec{c}$ is a vector generated from the sequence of the hidden states. f and q are some nonlinear functions. For instance, the Long-Short Term Memory (LSTM) is used as f, and $q(h_1, h_2, ..., h_{l-1}, h_l)$ as h_t.

Decoder

In RNN framework, the decoder is often used to predict the next word y_t given the context vector $\vec{c}$ and all the previously predicted words $y_1, y_2, ..., y_{t-1}$. In other words, the decoder defines a probability over the identification y_t by decomposing the joint probability into the ordered conditionals:

$$p(y) = \prod_{t=1}^{l} p(y_t | \{y_1, y_2, ..., y_{t-1}\}, \vec{c}) \tag{4}$$

4 Loanwords Identification

To make our loanwords identification model stronger, we also proposed several features as the additional knowledge of RNN. These features can be derived from monolingual corpus. In this section, we present two features firstly, then, we introduce the decoding part of our model.

4.1 Features

Inverse Language Model Feature

A language model is a probability distribution over sequences of words in NLP. Traditionally, language models are widely used in applications such as machine translation, speech recognition, part-of-speech tagging, parsing, and information retrieval. For example, in statistical machine translation, language models are used to improve the fluency of generated texts (Lembersky et al., 2012)

$$p(e|f) \propto p(f|e)p(e) \tag{5}$$

$p(e|f)$ indicates the translation model, which means the probability that the source string e is the translation of the target string f; $p(e)$ is the language model, which indicates the probability that the string e appeared in target language.

Usually, there are different pronunciation systems between donor language and recipient language. Different pronunciation rules can be represented by different character-based language models. In our paper, we propose an inverse language model (ILM) to constraint the output of loanword identification system.

N-gram Model

N-gram is a contiguous sequence of n items from a given text. An n-gram model models natural language sequences using the statistical properties of n-grams. Practically, an n-gram model predicts x_i based on $x_{i-(n-1)}, ..., x_{i-1}$. This can be indicated by probability terms as

$$p(x_i | x_{i-(n-1)}, ..., x_{i-1}) \tag{6}$$

When used in language modeling, independent assumptions are made so that each item word replies on its last $n-1$ words.

Inverse Language Model

As mentioned above, we can use a character-based language model to indicate a pronunciation system. Although a loanword may adapt the pronunciation of recipient language when borrowing, there are still some differences exist, and these differences usually reflect features of donor language pronunciation system. Accordingly, we computed the inverse language model feature as following

$$p_{ilm} = (1 - \lambda_1 p_{uyg}) + \lambda_2 p_{dnr} \tag{7}$$

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. Language model probabilities are all based on n-gram models.

Collocation Feature

Unlike Chinese, some words (e.g. person names) are written separately. For detect loanwords effectively, we proposed a collocation feature, which measures the co-occurrence probability of two words.

The frequency of words is a simple but effective metric in NLP. In this paper, we use a probability of words co-occurrence to measure the composition of several parts of a possible loanword. Similar to language model, we also use smooth mechanism to alleviate the data sparseness.

Another metric used in our collocation extraction is the skip-gram language model (SGLM), which has the ability to model semantic relations between words, and capture a form of compositionality. We represent words by distributed representation encoded in the hidden layers of neural networks. Given a word, the context can be learned by the model. In our model, we can apply the SGLM to predict a loanword (such as Chinese person names in Uyghur texts) based on one part of it.

$$p_{clc1} = \max \sum_{-k \le j-1, j \le k}^{l} \log p(w_{t+j}|w_t) \quad (8)$$

For example, if we indicate v as a function that maps one part of a loanword w to its n-dimensional vector representation, then

$$\begin{aligned} &v(\text{“}jinping\text{”}) - v(\text{“}shi\text{”}) + v(\text{“}zEmin\text{”}) \\ &\approx v(\text{“}jyang\text{”}) \end{aligned} \quad (9)$$

The symbol $\approx$ means its right hand side must be the nearest neighbor of the value of the left hand side.

4.2 Decoding

In this paper, we use a beam search decoder as the basic framework, a neural network and two features are also integrated into it. Two features used in this model are as two re-rankers to filter out incorrect outputs.

$$\begin{aligned} s(o|w) = &\log p_{rnn}(o|w) \\ &+ \mu_1 \log p_{ilm}(w_{c1...ck}) \\ &+ \mu_2 \log p_{clc}(w) \end{aligned} \quad (10)$$

Where μ_1 and μ_2 are parameters which determine how much inverse language model and collocation model are weighted. According to our model, several characteristics are captured when loanwords identification. $p_{rnn}(o|w)$ is the most important part in our framework, some information that difficult to define by human can be learned automatically by the RNN model. Different language has a different pronunciation system. In our loanwords identification task, we use the inverse language model of characters to highlight the probability of a loanword $p_{ilm}(w_{c1...ck})$. For loanwords such as person names which have been separated by blank spaces, we predict their contexts according to themselves. In our method, we integrated both inner word information and information between words into the decoder.

5 Experiments and Results

We conduct experiments to evaluate our loanword identification model. According to the tasks defined in this paper, these experiments can be divided into two types: 1) in-domain experiments; 2) cross-domain experiments. We train the loanword identification model using a small set of training data, and evaluate the performance of our model with three held out test sets for each language.

5.1 Setup and Datasets

We evaluate our approach on three donor languages: Chinese, Russian and Arabic. In our approach, loanword identification models are trained on Uyghur news corpora. Test sets used in our experiments include in-domain test sets and cross-domain test sets. Since we are very familiar with Chinese and Russian, we labeled several types of Chinese (Chn) and Russian (Rus) loanwords in Uyghur test sets, such as person names, locations, and other daily used words. For Arabic loanwords (Arab), we labeled them in test set manually. Because we have limited knowledge about Arabic, we just labeled some person names and locations. We collected some relatively regular corpora from news websites in Chinese[2] , Russian[3] and Arabic[4] to train the language models, respectively.

We built the recurrent neural network which used in our loanword identification model on the open source deep learning software Deeplearning4j[5]. For

[2]http://www.people.com.cn/
[3]http://sputniknews.ru/
[4]http://arabic-media.com/arabicnews.htm
[5]http://deeplearning4j.org

Languages	TR-Set	DE-Set	TE-Set
Uyghur	10,000 * 3	1,000 * 3	1,000 * 3
Chinese	\	1,000	\
Russian	\	1,000	\
Arabic	\	1,000	\

Table 2: Statistic of Corpora.

the inverse language model, we used a java version language model tool which was implemented by ourselves. For the collocation extraction feature, we trained a model based on the word2vec, which was proposed by Tomas Mikolov, and a java version is also implemented in Deeplearning4j toolkits.

To evaluate the performance of loanword identification models, several metrics are used in our experiments:

$$R = \frac{A}{A+C}, P = \frac{A}{A+B}, F1 = \frac{2 \star R \star P}{P+R} \quad (11)$$

$P(Precision)$ indicates the percentage of loanwords found that match exactly the spans found in the evaluations data (test set);

$R(Recall)$ means the percentage of loanwords defined in the corpus that were found in the same location;

$F1$ can be interpreted as the harmonic mean of P and R.

5.2 Experiments and Analysis

For validate the effectiveness of our loanword identification model, we first compare our model (RNN-based model) with other loanword detection models, including CRFs-based model (CRFs) (Lafferty et al., 2001), the identification model based on string similarity (SSIM) (Mi et al., 2013), classification-based identification model (CBIM) (Mi et al., 2014). We suggest two important features in this paper to optimize the output of our loanword identification model, affection of these features are evaluated on identification performance. Loanwords can exist in any domains in a language; therefore, we also conduct experiments on texts in several domains.

Evaluation on Loanword Identification Models

In this part, we introduce the experiment results on four models, then we analysis the reasons.

From the Table 3 we can found that the performance of RNN based model outperforms other three approaches, we summarized possible reasons as follow: 1) CRFs model rely heavily on labeled data, because we only have limited training examples, the CRFs model achieved lowest performance among four models; 2) SSIM model based on two string similarity algorithms: edit distance and the common substring, compare with the RNN model, SSIM has a limited ability of generalization, and cannot capture semantic information in Uyghur texts, so the SSIM achieved a relative low performance; 3) Several information including above two algorithms are integrated into the CIBM model, and consider the loanwords identification as a classification problem, the performance of CIBM model outperforms the CRFs model and SSIM model. However, like the SSIM model, there is almost no semantic information and limited generalization ability, therefore, the performance of CIBM model cannot achieve or outperform the RNN based model.

Evaluation on Features Used in RNN-based Model

Features used in our model optimized the output of loanword identification. We show the experimental results on combination of features: $RNN+f0$ (no additional feature used), $RNN+f1$ (inverse language model feature used), $RNN+f2$ (collocation feature used) and $RNN+f1+f2$ (both inverse language model feature and collocation feature are used).

In Table 4, $RNN+f1$ combines the inverse language model information into loanword identification model, which apply the local feature in our task, so the performance of $RNN+f1$ outperforms the basic RNN model and $RNN+f2$. $RNN+f2$ integrated the collocation information into the model, and the generation ability of the model only rely on RNN, therefore, the performance of $RNN+f2$ only outperform the basic RNN model. The $RNN+f1+f2$ not only combine the generalization ability into the model, but also the local feature ($f1$) and global feature ($f2$). So

Languages	P-Chn	R-Chn	F1-Chn	P-Rus	R-Rus	F1-Rus	P-Arab	R-Arab	F1-Arab
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
RNNs	**78.97**	**79.20**	**79.08**	**82.55**	**75.93**	**79.10**	**83.26**	**77.58**	**80.32**

Table 3: Experimental Results on Loanword Identification Models.

Languages	P-Chn	R-Chn	F1-Chn	P-Rus	R-Rus	F1-Rus	P-Arab	R-Arab	F1-Arab
RNN+f0	77.65	67.89	72.44	78.02	68.33	72.85	78.38	70.96	74.49
RNN+f1	78.86	70.32	74.35	81.94	70.65	75.88	81.12	71.52	76.02
CIRNN+f2	78.79	69.54	73.88	81.35	71.28	75.98	80.76	70.20	75.11
RNN+f1+f2	**78.97**	**79.20**	**79.08**	**82.55**	**75.93**	**79.10**	**83.26**	**77.58**	**80.32**

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

the $RNN + f1 + f2$ model achieved the best performance.

Evaluation on Cross-domain Corpora

We evaluate our model in two domains on different test sets: $RNNLIS + NEWS$ and $RNNLIS + ORAL$.

In Table 5, the experimental results on news ($RNNLIS + NEWS$) which is similar with our training examples are outperform the results on oral test set ($RNNLIS + ORAL$). This may be no doubt. Amazingly, we found that performance of $RNNLIS + ORAL$ is just a little worse compared with $RNNLIS + NEWS$. A possible reason is that our model can learn representation of knowledge beyond given training examples.

5.3 Discussion

In our experiments, we try to identify Chinese, Russian and Arabic loanwords in Uyghur texts. We found that results on Arabic loanwords identification achieved the best performance. There are two possible reasons. First, most of loanwords labeled in the training examples for Arabic loanwords identification are person names; therefore, it is relatively easy to find them out. Second, Persian has exerted some influence on Arabic, and borrowing much vocabulary from it. Meanwhile, Uyghur has historically been influenced strongly by Persian, so Arabic loanwords in Uyghur may have the similar pronunciation system with Arabic. These two reasons contribute to Arabic loanwords identification in Uyghur.

We have limited number of labeled corpus, so a competitive identification result cannot be expected if a traditional approach is used (such as the CRF). Our proposed RNN Encoder-Decoder framework can learn features automatically and use its internal memory to process arbitrary sequences of inputs. Additionally, two features inverse language model and collocation can constraint the output of identification model. Therefore, our model achieved the best performance.

Loanword identification models are all trained on news corpora, so in cross-domain (news and oral) experiments, results in news are outperform results in oral. We analysis the results, and found that several errors including spelling error are exist in oral corpora, and these errors may affect the performance of our model.

6 Related work

There has been relatively few previous works on loanwords identification in Uyghur. Our work is inspired by two lines of research: (1) recurrent neural network; (2) loanwords detection.

6.1 Recurrent Neural Network

In recent years, the Recurrent Neural Network has proven to be highly successful in capturing semantic information in text and has improved the results of several tasks in NLP area. (Socher et al., 2013) uses

Languages	P-Chn	R-Chn	F1-Chn	P-Rus	R-Rus	F1-Rus	P-Arab	R-Arab	F1-Arab
RNNLIS+NEWS	**78.97**	**79.20**	**79.08**	**82.55**	**75.93**	**79.10**	**83.26**	**77.58**	**80.32**
RNNLIS+ORAL	75.23	76.44	75.83	78.11	70.59	74.16	80.03	76.42	78.18

Table 5: Evaluation on Cross-Domain Corpora.

a recursive neural network to predict sentence sentiment. (Luong et al., 2013) generates better word representation with recursive neural network. (Cho et al., 2014a) proposed a RNN encoder-decoder model to learn phrase representations in SMT. (Irsoy et al., 2014) introduce a deep recursive neural network, and evaluate this model on the task of fine-grained sentiment classification. (Liu et al., 2014) propose a recursive recurrent neural network to model the end-to-end decoding process for SMT; experiments show that this approach can outperform the state-of-the-art baseline. (Yao et al., 2013) optimized the recurrent neural network language model to perform language understanding. (Graves , 2012) apply a RNN based system in probabilistic sequence transduction.

6.2 Loanwords Detection

In general, word borrowing is often concerned by linguists (Chen, 2011; Chen et al., 2011a). There are relatively few researches about loanwords in NLP area. (Tsvetkov et al., 2015a) and (Tsvetkov et al., 2016) proposed a morph-phonological 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 et al., 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., 2013). (Mi et al., 2014) propose a loanword detection method based on the perceptron model, several features are used in model training.

7 Conclusion

We have presented an approach to identify loanwords (Chinese, Russian and Arabic loanwords) in Uyghur texts, our model based on the RNN Encoder-Decoder framework. We also suggested two important features: inverse language model and collocation feature to optimize the output of our loanword identification model. Experimental results show that our model achieves significant improvements in loanwords detection tasks. In the future, we plan to further validate the effectiveness of our approach on more languages, especially on languages with rich morphology.

Acknowledgments

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 and the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No.XDA06030400.

References

Shiming Chen. 2011. New Research on Chinese Loanwords in the Uyghur Language. *N.W.Journal of Ethnology,* pages 176-180, 28(1).

Yan Chen and Ping Chen. 2011. A Comparison on the methods of Uyghur and Chinese Loan Words. *Journal of Kashgar Teachers College* pages 51-55, 32(2).

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8),* pages 103-111, October 25, Doha, Qatar. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder ‐ Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014),* pages 1724-1734, October 25-29, Doha, Qatar. Association for Computational Linguistics.

Alex Graves. 2012. Sequence Transduction with Recurrent Neural Networks. In *Proceedings ICML Representation Learning Workshop,* Edinburgh, Scotland.

Ozan Irsoy and Claire Cardie. 2012. Deep Recursive Neural Networks for Compositionality in Language. In *Proceedings of the 2014 Conference on Advances in Neural Information Processing Systems (NIPS 2014),* pages 2096-2104, December 8-13, Montréal, Canada.

Herbert Jaeger. 2002. A tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the "echo state network" approach. *GMD-Forschungszentrum Informationstechnik*.

John Lafferty, Andrew McCallum and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001),* pages 282-289, June 28-July 2, Bellevue, Washington, USA.

Gennadi Lembersky, Noam Ordan and Shuly Wintner. 2012. Language Models for Machine Translation: Original vs. Translated Texts. *Computational Linguistics,* pages 799-825, 38(4). Association for Computational Linguistics.

Shujie Liu, Nan Yang, Mu Li and Ming Zhou. 2014. A Recursive Recurrent Neural Network for Statistical Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014),* pages 1491-1500, June 23-25, Baltimore, Maryland, USA. Association for Computational Linguistics.

Thang Luong, Richard Socher and Christopher Manning. 2013. Better Word Representations with Recursive Neural Networks for Morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning (CoNLL 2013),* pages 104-113, August 8-9, Sofia, Bulgaria. Association for Computational Linguistics.

Chenggang Mi, Yating Yang, Xi Zhou, Xiao Li and Mingzhong Yang. 2013. Recognition of Chinese Loan Words in Uyghur Based on String Similarity. *Journal of Chinese Information Processing,* pages 173-179, 27(5).

Chenggang Mi, Yating Yang, Lei Wang, Xiao Li and Kamali Dalielihan. 2014. Detection of Loan Words in Uyghur Texts. In *Proceedings of the 3rd International Conference on Natural Language Processing and Chinese Computing (NLPCC 2014),* pages 103-112, December 5-9, Shen Zhen, China.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL (COLING-ACL 2006),* pages 81-88, July 17-21, Sydney, Australia. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013),* pages 1631-1642, October 18-21, Seattle, Washington, USA. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 2014 Conference on Advances in Neural Information Processing Systems (NIPS 2014),* pages 3104-3112, December 8-13, Montréal, Canada.

Yulia Tsvetkov, Waleed Ammar and Chris Dyer. 2015. Constraint-Based Models of Lexical Borrowing. In *Proceedings of the 2015 Conference on Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL (NAACL-HLT 2015),* pages 598-608, May 31-June 5, Denver, Colorado.

Yulia Tsvetkov and Chris Dyer. 2016. Cross-Lingual Bridges with Models of Lexical Borrowing. *Journal of Artificial Intelligence Research* pages 63-93, 55(2016).

Yulia Tsvetkov and Chris Dyer. 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 (Short Papers)(ACL-IJCNLP 2015),* pages 125 – 131, July 26-31, Beijing, China.

Kaisheng Yao, Geoffrey Zweig, Mei-Yuh Hwang, Yangyang Shi and Dong Yu. 2013. Recurrent Neural Networks for Language Understanding. In *Proceedings of 14th Annual Conference of the Inter national Speech Communication Association (INTERSPEECH 2013),* pages 2524-2528, August 25-29, Lyon, France.