Learning Distributed Representations of Uyghur Words and Morphemes

Halidanmu Abudukelimu[1], Yang Liu[1,2], Xinxiong Chen[1], Maosong Sun[1,2] and Abudoukelimu Abulizi[3]
[1] State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Sci. and Tech., Tsinghua University, Beijing, China
[2]Jiangsu Collaborative Innovation Center for Language Competence, Jiangsu, China
[3]Lab of Computational Linguistics, Center for Psychology and Cognitive Science
School of Humanities, Tsinghua University Beijing, China
abdklmhldm@gmail.com, {liuyang2011,sms}@tsinghua.edu.cn,
cxx.thu@gmail.com, keram1106@163.com

Abstract. While distributed representations have proven to be very successful in a variety of NLP tasks, learning distributed representations for agglutinative languages such as Uyghur still faces a major challenge: most words are composed of many morphemes and occur only once on the training data. To address the data sparsity problem, we propose an approach to learn distributed representations of Uyghur words and morphemes from unlabeled data. The central idea is to treat morphemes rather than words as the basic unit of representation learning. We annotate a Uyghur word similarity dataset and show that our approach achieves significant improvements over CBOW, a state-of-the-art model for computing vector representations of words.

Keywords: distributed representations; Uyghur; word; morpheme

1 Introduction

Developing natural language processing techniques for Uyghur is difficult, not only because of the unavailability of publicly accessible annotated corpora, but also due to its nature of agglutination. On one hand, the annotated corpora of Uyghur for morphological analysis, POS tagging, parsing, translation and sentiment analysis are far more limited in both quantity and coverage as compared with resource-rich languages such as English and Chinese. On the other hand, Uyghur words often consist of many morphemes and differ significantly from English and Chinese in terms of morphology and syntax, making it difficult to directly adopt state-of-the-art NLP models and algorithms.

Fortunately, unsupervised learning of distributed representations brings hope to addressing the resource scarcity problem. In recent years, learning distributed representations of words from unlabeled data has received intensive attention [1,6,9]. Distributed representations, which are continuous dense real-valued vectors, are capable of capturing multiple degrees of syntactic and semantic similarities between words.

UYGHUR PRODUCTIVE DERIVATION	
يەر yer	land
يەرلىك yerlik	local
يەرلىكلەش yerliklex	to be localed
يەرلىكلەشتۈر yerliklextür	localized
يەرلىكلەشتۈرۈل yerliklextürül	to be localized
يەرلىكلەشتۈرۈلمە yerliklextürülme	to not be localized
يەرلىكلەشتۈرۈلمەيمىز yerliklextürülmeymiz	We unable to be localized

Fig. 1. Example: Uyghur words and their corresponding English translations.

These representations have proven to benefit many NLP tasks including language modeling [1,11], machine translation [2], and semantic analysis [7].

However, most existing methods treat words as the atomic units in distributed representation learning [1,6,9]. This is problematic for agglutinative languages such as Uyghur in which most words are composed of many morphemes. As most Uyghur words only occur once on the training data, it is hard for approaches treating words as the basic unit to learn vector representations accurately due to the data sparsity. To address this problem, a number of authors propose to learn word presentations by exploiting the minimum meaning bearing units such as characters in Chinese and morphemes in Russia [8,2,12,3].

In this work, we follow this line of research to learn distributed representations of Uyghur words and morphemes from unlabeled data. The basic idea is to treat Uyghur morphemes as the atomic unit to account for the internal structure of Uyghur words. We propose a morpheme-enhanced continuous bag-of-words (mCBOW) model that uses morpheme vectors to derive word vectors. We annotate a Uyghur word similarity dataset and show that our approach achieves significant improvements over CBOW [9], a state-of-the-art model for computing vector representations of words.

2 Background

Uyghur belongs to the Karluk branch of the Turkic language family and is spoken mainly by the Uyghur people in the Xinjiang Uyghur Autonomous Region of Western China. Similar to many other Turkic languages, Uyghur is agglutinative, lacks grammatical articles and noun classes. The basic word order of Uyghur is subject-object-verb.

Fig. 1 shows some Uyghur words and their corresponding English translations. One single Uyghur word usually contains rich information by combining various morphemes including stems, prefixes, and affixes.

Due to the scarcity of resources for Uyghur processing, it is appealing to learn distributed representations of Uyghur words and morphemes from unlabeled data using the continuous bag-of-words (CBOW) model [9]. The intuition is that a good model should be able to predict a word given its surrounding context.

Fig. 2(a) illustrates the idea of CBOW. Given a Uyghur sentence *ular mekteptin kaldi*, the model aims to predict *mekteptin* given the context words *ular* and *kaldi*, which are all represented as real-valued vectors. These distributed representations are surprisingly good at capturing syntactic and semantic regularities in language [9].

More formally, given a training corpus $D = \{w_1, \ldots, w_T\}$, the training objective of CBOW is to maximize the average log probability

$$\frac{1}{T}\sum_{t=1}^{T} \log P(w_t|w_{t-c}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+c}) \tag{1}$$

where c is the size of the training context around the center word w_t. The prediction probability can be defined using a softmax function

$$P(w_t|w_{t-c}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+c}) = \frac{\exp(\mathbf{u}_{context}^{\top} \cdot \mathbf{v}_{w_t})}{\sum_{w' \in W} \exp(\mathbf{u}_{context}^{\top} \cdot \mathbf{v}_{w'})} \tag{2}$$

where W is the vocabulary, $\mathbf{v}_{w_t}$ is the input vector of w_t, and $\mathbf{u}_{context}^{\top}$ is the output vector of the surrounding context:

$$\mathbf{u}_{context}^{\top} = \frac{1}{2c} \sum_{t-c \le i \le t+c, i \ne t} \mathbf{u}_{w_i} \tag{3}$$

Note that the output vector of the surrounding context is the average of all context word vectors.

Although the CBOW model works well for many languages such as English and Chinese, it faces a severe data sparsity problem when processing agglutinative languages such as Uyghur: most words only occur once on the training data. As a result, modeling at the word level is insufficient to capture the linguistic regularities in morphologically-rich languages.

3 Morpheme-Enhanced CBOW

A number of authors have proposed to exploit the internal structures of words to address the data sparsity problem [8,2,12,3]. The central idea is that the minimum meaning-bearing units, say morphemes in Uyghur or characters in Chinese, are also modeled as real-valued vectors of parameters to derive the vectors of words. While Luong et al. [8] leverage recursive neural networks to model the internal hierarchical structure, Botha and Blunsom [2] and Chen et al. [3] simply use addition as composition function. They

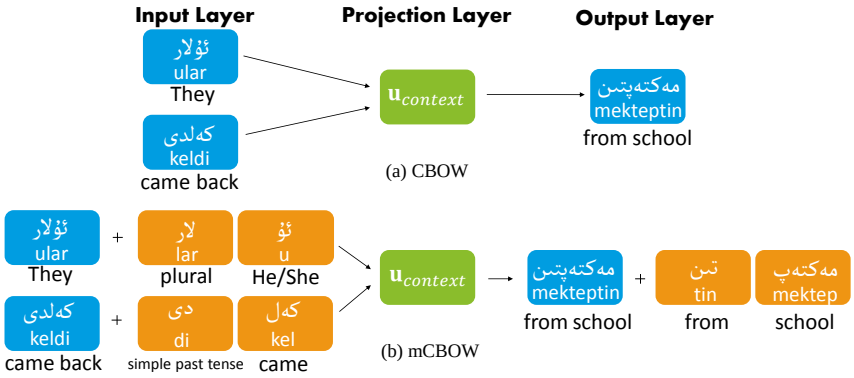

Fig. 2. (a) The continuous bag-of-words (CBOW) model and (b) the morpheme-enhanced continuous bag-of-words (mCBOW) model. Given a Uyghur sentence *ular mekteptin kaldi*, the mCBOW model predicts the word *mekteptin* by taking two context words *ular* and *kaldi* as input. mCBOW differs from CBOW in that it takes both word and morpheme vectors into account.

show that modeling at the morpheme or character level outperforms modeling at the word level for a variety of languages.

In this work, we follow this line of research and propose a **morpheme-enhanced continuous bag-of-words (mCBOW)** model for Uyghur. As shown in Fig. 2(b), mCBOW extends CBOW to consider both word and morpheme vectors, highlighted in blue and yellow, respectively. The basic idea is to derive word vectors from morphme vectors:

$$\mathbf{v}_{\text{unfortunately}} = \mathbf{v}_{\text{un}} + \mathbf{v}_{\text{fortunate}} + \mathbf{v}_{\text{ly}}$$

We hope that the inclusion of morpheme vectors enables the model to be more robust to data sparsity.

More formally, suppose a Uyghur word w is composed of K morphemes: $w = m_1, \ldots, m_K$. We use m_k to denote the k-th morpheme in the word. Following Botha and Blunsom [2], the vector representation of w can be computed using the vectors of morphemes:

$$\tilde{\mathbf{u}}_w = \mathbf{u}_w + \sum_{k=1}^{K} \mathbf{u}_{m_k} \tag{4}$$

$$\tilde{\mathbf{v}}_w = \mathbf{v}_w + \sum_{k=1}^{K} \mathbf{v}_{m_k} \tag{5}$$

Note that the surface form of a word (i.e., $\mathbf{u}_w$ and $\mathbf{v}_w$) is also included as a factor to account for noncompositional constructions as suggested by Botha and Blunsom [2]. They indicate that this strategy also overcomes the order-invariance of additive composition.

Given a training corpus $D = \{w_1, \ldots, w_T\}$, the training objective of mCBOW is still to maximize the average log probability as shown in Eq. (1). The prediction

probability can be defined using a softmax function

$$P(w_t|w_{t-c},\ldots,w_{t-1},w_{t+1},\ldots,w_{t+c}) = \frac{\exp(\tilde{\mathbf{u}}_{context}^{\top}\cdot\tilde{\mathbf{v}}_{w_t})}{\sum_{w'\in W}\exp(\tilde{\mathbf{u}}_{context}^{\top}\cdot\tilde{\mathbf{v}}_{w'})} \tag{6}$$

where $\tilde{\mathbf{v}}_{w_t}$ is the input vector of w_t, and $\tilde{\mathbf{u}}_{context}^{\top}$ is the output vector of the surrounding context:

$$\tilde{\mathbf{u}}_{context}^{\top} = \frac{1}{2c}\sum_{t-c\leq i\leq t+c, i\neq t}\tilde{\mathbf{u}}_{w_i} \tag{7}$$

Following Mikolov et al. [9], we exploit stochastic gradient descent (SGD) and negative sampling to train the mCBOW model. The gradients are calculated using the back-propagation algorithm. The word and morpheme vectors are initialized randomly.

It is clear that the mCBOW model is a natural extension of CBOW that includes the vectors of morphemes. As we use addition as the composition function, it is still easy and efficient to train the model using negative sampling [9]. Our work is also in spirit close to Botha and Blunsom [2] and Chen et al. [3]. While both Botha and Blunsom [2] and mCBOW use addition as composition function at the morpheme level, Botha and Blunsom [2] exploit the log-bilingual model [10] at the word level but we leverage the CBOW model. The difference from Chen et al. [3] is that our model considers the morphemes of the word to be predicted, which is very useful for improving the accuracy.

4 Experiments

We evaluate our approach on a Uyghur word similarity task. We build a Uyghur word similarity dataset, which we refer to as *uyWordSim-353*, by manually translating the popular *WordSim-353* [5] into Uyghur. The training set contains news articles from the Tianshan website with 1.26M words. [1] We use Morfessor v0.9.2 [4] to segment Uyghur words into morphemes by setting the parameter "PPLTHRESH" to 200. The evaluation metric is Spearman's rank correlation coefficient ($\rho \times 100$) between similarity scores assigned by the model and by human annotators.

4.1 Data Sparsity in Uyghur

We find on the training data that the average lengths of words, stems, affixes, and suffixes are 17, 14, 5, and 4, respectively. The maximum length of a Uyghur word is 33 characters. 53% of words contain at least two morphemes.

As shown in Table 1, over 80% of words occur no greater than 10 times on the training data. In particular, about 47% of words occur only once. This leads to severe data sparsity for learning distributed representations of Uyghur words.

4.2 Comparison with CBOW and Skip-Gram

We compare mCBOW with CBOW and Skip-Gram. The Skip-Gram model is a reverse variant of CBOW: predicting the surrounding context given a specific word [9]. We

[1] http://uy.ts.cn

Freq.	# Words	Percent. (%)
1	32,457	47.66
2	9,524	13.98
3	4,691	6.89
4	3,045	4.47
5	2,199	3.23
6	1,617	2.37
7	1,223	1.80
8	974	1.43
9	843	1.24
10	717	1.05
> 10	10,812	15.88
Total	68,102	100.00

Table 1. The distribution of word frequencies. We find that 47.66% of Uyghur words occur only once on the training data.

Method	*uyWordSim*	
	353 pairs	196 pairs
CBOW	8.21	43.12
Skip-Gram	9.22	45.34
mCBOW	**10.88**	**45.50**

Table 2. Comparison with CBOW and Skip-Gram. "353 pairs" denotes the the Uyghur translations of the original *WordSim-353* dataset. "196 pairs" denotes a subset of "353 pairs" that removes words that do not occur on the training data.

set the number of negative examples in negative sampling to 10 and run the training algorihm for 30 iterations.

We find that there are many words in *uyWordSim-353* do not occur on the training data. In addition, the Uyghur translation of an English word is sometimes a phrase. To handle these OOV words and phrases, the similarity scores of these OOV words are set to -1. After removing these OOV words and phrases, we obtain a subset called *uyWordSim-196*.

As shown in Table 2, mCBOW significantly improves over CBOW and Skip-Gram. Note that the correlation coefficients are very low due to the presence of OOV words and phrases as well as the severe data sparsity. Our approach also outperforms CBOW and Skip-Gram on the *uyWordSim-196* dataset, in which all words occur on the training data. mCBOW still achieves higher accuracy than CBOW, suggesting that modeling the internal structures of Uyghur words does benefit representation learning.

The improvement over Skip-Gram on the *uyWordSim-196* dataset is insignificant because Skip-Gram itself is better than CBOW on this task. It is possible to extend our approach to morpheme-enhanced Skip-Gram. We leave this for future work.

Freq.	Pairs	CBOW	mCBOW
> 100	137	46.96	44.01
< 100	59	34.76	51.98
< 50	50	35.17	48.62
< 30	29	19.70	24.14
< 20	18	-0.49	36.52
< 10	14	6.59	26.92

Table 3. Effect of word frequencies on accuracy. While CBOW works well for high-frequency words, mCBOW is more capable of handling infrequent words.

# Morph.	Pairs	CBOW	mCBOW
1	167	48.12	47.83
> 1	30	5.52	26.23

Table 4. Effect of morpheme count on accuracy. "# Morph." denotes the number of morphemes in a Uyghur word. While 53% of words contain at least two morphemes on the training data, the percentage is 85% on the *uyWordSim-196* dataset. mCBOW outperforms CBOW when dealing with multi-morpheme Uyghur words.

4.3 Effect of Word Frequencies

We find most words in the *uyWordSim-196* dataset occur more than 100 times on the training data. To investigate the effect of word frequencies on the accuracy, we compare CBOW and mCBOW on various subsets of *uyWordSim-196* in terms of word frequencies.

As shown in Table 3, CBOW achieves a higher accuracy than mCBOW on 137 word pairs that occur more than 100 times on the training data, indicating that it is unnecessary to consider internal structures of words if the training data is not sparse. However, the accuracy of CBOW drops dramatically with the decrease of word frequencies and even achieves a negative correlation coefficient. In contrast, our approach is more robust to data sparsity.

4.4 Effect of Morpheme Count

On the training data, 53% of words contain at least two morphemes and 47% of words occur only once. However, this is not the case on the test set because 85% of Uyghur words only contain stems.

Table 4 shows the effect of morpheme count on accuracy. If a Uyghur word only contains stem, CBOW slightly outperforms mCBOW but the difference is not significant. Dealing with multi-morpheme Uyghur words, however, mCBOW improves over CBOW by a large margin. This finding confirms the effectiveness of our approach.

4.5 Case Study

Fig. 3 shows the top 5 closest words to some Uyghur infrequent words obtained from CBOW and mCBOW. It is clear that our approach is capable of capturing the semantic similarities between Uyghur words.

Words	CBOW	mCBOW
ۋيېتنام Vietnam	ساغلاملىقىغا health	ئەرەبىستان Saudi Arabia
	ئۆكتەبىرنى October	ببرنيې (place name)
	سەھنىلەرگە stage	ۋېنگرىيە Hungary
	شەنرىن (person name)	جۇڭگو China
	لوڭقىسى (World) Cup	راللى (a kind of sport game)
يۇمۇر humor	ساۋاتلىرىمۇ common sense	شېئىر peom
	سۇنۇشنىلا present	ئەدەبيې literature
	ئاتكېرتكا greeting card	ئەتكەن make
	شېئىر poem	نادىر excellent
	سۆزلىيەلەيدۇ can speak	يازالايدىغان can write
جىسىم entity	كونتورى contour	ئاقار meteor
	فازىلىق phase	ئۇلترا infrared
	مۇڭغۇلچە Mongolian	گۈلنى flower
	ھىدروگېن hydrogen	ئاكۇپ tunnel
	سانو (person name)	جىسمانيې materially

Fig. 3. Top 5 nearest neighbors of example Uyghur words.

As shown in the figure, the results of the nearest neighbors of *Vietnam* returned by mCBOW are better than CBOW. In mCBOW, almost all the nearest words are semantically closely related to *Vietnam* except *(a kind of sport game)*. However, in CBOW, the results differ a great deal from the semantic meaning of Vietnam. The results for *humor* and *entity* are similar as well (In CBOW, *common sense*, *present*, *greeting card* and *can speak* are unrelated to *humor*, *contour*, *phase*, *Mongolian* and *personal name* have little correlation with *entity*).

5 Related Work

Distributed word representations, low dimension real-valued vectors for words, usually capturing both semantic and syntactic information of words. These representations have been successfully used in a variety of NLP tasks. Most word representation models exhibits high computational complexity, which makes them unable to work for large-scale text corpora efficiently. Recently, Mikolov et al. [9] proposed two efficient models, continuous bag-of-words mode (CBOW) and Skip-Gram model, to learn word embeddings from large-scale text corpora. The training objective of CBOW is to combine the embeddings of context words to predict the target word; while Skip-Gram is to use the embedding of each target word to predict its context words.

The unsupervised learning of distributed representations on large corpus brings hope to addressing the resource scarcity problem of Uyghur. However, most existing methods treat words as the atomic units in distributed representation learning [1,6,9].

This is problematic for agglutinative languages such as Uyghur in which most words are composed of many morphemes. To address this problem, a number of authors propose to learn word presentations by exploiting the minimum meaning bearing units.

To learn morpheme representations, Lazaridou et al. [7] had used compositional distributional semantic models, originally designed to learn meanings of phrases, to derive representations for complex words. Luong et al. [8] also choose to operate at the morpheme level and used a recursive neural network to explicitly model the morphological structures of words and learn morphologically-aware embeddings. Botha and Blunsom [2] used addition as composition function at the morpheme level and exploit the log-bilingual model. Chen et al. [3] decided to learn representations at the character level and proposed multiple-prototype character representations to deal with the ambiguity problem of characters.

In this paper, we focus on Uyghur and follow this line of work to learn both word and morpheme representations of Uyghur.

6 Conclusion

We have presented a morpheme-enhanced continuous bag-of-words (mCBOW) model for learning vector representations of Uyghur words and vectors. The model treats morphemes as the basic unit and uses addition as the composition function. Experiments on the Uyghur word similarity task show that our approach significantly outperforms the CBOW model. In particular, the mCBOW model is more capable of handling infrequent and multi-morpheme Uyghur words than CBOW.

Note that our model is an unsupervised model, thus it can also be applied to other agglunative languages, like: Turkish, Uzbek, Kazak, etc.

In the future, we plan to apply our idea to more models such as the log-bilingual model [10] and the Skip-Gram model [9]. It is also interesting to model the recursive structure of Uyghur words like Luong et al. [8].

Acknowledgments

This research is supported by National Key Basic Research Program of China (973 Program 2014CB340500), the National Natural Science Foundation of China (No. 61331013), the National Key Technology R & D Program (No. 2014BAK10B03), the Singapore National Research Foundation under its International Research Center @ Singapore Funding Initiative and administered by the IDM Programme. We are grateful to Meiping Dong, Lei Xu, Liner Yang, Yu Zhao, Yankai Lin, Chunyang Liu, Shiqi Shen, and Meng Zhang for their constructive feedback to the early draft of this paper.

References

1. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of Machine Learning Research (2003)
2. Botha, J.A., Blunsom, P.: Compositional morphology for word representations and language modelling. In: Proceedings of ICML (2014)

3. Chen, X., Xu, L., Liu, Z., Sun, M., Luan, H.: Joint learning of character and word embeddings. In: Proceedings of IJCAI (2015)
4. Creutz, M., Lagus, K.: Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech Language Processing (2007)
5. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Sloan, Z., Wolfman, G., Ruppin, E.: Placing search in context: the concepted revisited. ACM Transactions on Information Systems 20 (2002)
6. Huang, E., Socher, R., Manning, C.D., Ng, A.Y.: Improving word representations via global context and multiple word prototypes. In: Proceedings of ACL (2012)
7. Lazaridou, A., Marelli, M., Zamparelli, R., Baroni, M.: Compositionally derived representations of morphologically complex words in distributional semantics. In: Proceedings of ACL (2013)
8. Luong, M.T., Socher, R., Manning, C.D.: Better word representations with recursive neural networks for morphology. In: Proceedings of CoNLL (2013)
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of NIPS (2013)
10. Mnih, A., Hinton, G.: Three new graphical models for statistical language modelling. In: Proceedings of ICML (2007)
11. Mnih, A., Hinton, G.: A scalable hierarchical distributed language model. In: Proceedings of NIPS (2008)
12. Qiu, S., Cui, Q., Bian, J., Gao, B., Liu, T.Y.: Co-learning of word representations and morpheme representations. In: Proceedings of COLING (2014)