Construction of an English-Uyghur WordNet Dataset

Kahaerjiang Abiderexiti and Maosong Sun *

Department of Computer Science and Technology
Institute for Artificial Intelligence
State Key Lab on Intelligent Technology and Systems
Tsinghua University, Beijing, China
khejabdr15@mails.tsinghua.edu.cn, sms@tsinghua.edu.cn

Abstract. Automatically building semantic resources is essential to low resource-languages like Uyghur. However, Uyghur suffers from a lack of publicly available evaluation dataset for automatically building semantic resources like WordNet. To cope with this problem, first, we build the largest Uyghur-English and English-Uyghur dictionaries by exploiting many possible online and offline resources. Then by using Princeton WordNet (PWN)3.0 and Contemporary Uyghur Detailed Dictionary (CUDD), we construct an English-Uyghur WordNet evaluation dataset which is publicly available (https://github.com/kaharjan/uywordnet). In this dataset, more than 73,000 English synsets are mapped Uyghur automatically, in which over 20,000 are annotated manually. And the corresponding Uyghur words include definition and examples in Uyghur language context. We also propose a Synset Mapping based on Word Embeddings(SMWE) method. The experimental results on the dataset are promising.

Keywords: Uyghur · WordNet · Dataset · Synset mapping .

1 Introduction

Since the introduction of Princeton WordNet (PWN) by professor George A. Miller [10], the construction of WordNet in other languages has begun. So far, WordNets in more than 50 languages have been constructed [1]. However, little research has been done on Uyghur WordNet construction or evaluation. Although BabelNet[2] contains Uyghur, its Uyghur parts are based on Wikipedia in Uyghur which contains around 2000 articles, and most of them are not aligned to other languages. So aligned concepts in Uyghur in the knowledge graph are scarce, and there are few definitions of these concepts in Uyghur. English-Uyghur WordNet constructed merely by translating PWN would suffer from lacking definitions and example sentences in Uyghur context. This would narrow the

* Corresponding author: M. Sun (sms@tsinghua.edu.cn)

[1] http://globalwordnet.org/resources/wordnets-in-the-world/

[2] http://babelnet.org/

value of Uyghur WordNet to a certain extent and help downstream applications little. Also, there are no off-the-shelf English-Uyghur or Uyghur-English dictionaries. Further, there is no evaluation dataset for automatic construction of English-Uyghur WordNet. As a result, there is little research about the automatic construction of English-Uyghur WordNet. Fortunately, Contemporary Uyghur Detailed Dictionary (CUDD) is available. The dictionary contains more than 60,000 words. Each sense of each word is explained by a definition and some example sentences. So far,only Aierken *et al.* [6] attempted to construct a semantic lexicon using CUDD. Since then, the dictionary has not been used for WordNet construction.

To address those problems, first, we preprocess CUDD to get all words' POS tags, definitions, and example sentences. We also build the largest English-Uyghur and Uyghur-English dictionary from online and offline using most of the available resources. Then, we map PWN 3.0 synsets to CUDD nouns using the dictionary to get a preliminary mapping dataset. And we manually annotate more than 20,000 English synsets to Uyghur to get high-quality development and test dataset. Finally, by coping with the common problems of Uyghur language processing like stemming and word representation, we propose a synset mapping method based on word embeddings using our dataset. The experimental results are promising. We make our dataset available online. In our dataset, every Uyghur word has a definition in Uyghur and some example sentences in Uyghur language context which would improve the value of this dataset to the downstream applications.

2 Related Work

In more than 30 years of the WordNet construction in various languages, many construction methods are proposed. The construction methods of WordNet can be divided into the following three kinds: construction by humans, external resource linking, and automatically.

PWN is built manually [10]. Wang *et al.* [25] merge the Southeast University WordNet (SEW)[26], Sinica Bilingual Ontological WordNet (Sinica BOW) [12], China Taiwan University WordNet (CWN) [13] and the Chinese part of the open multilingual WordNet (OMW) [8] , then manually proofread to construct Chinese Open WordNet (COW). For the external resource linking methods, through a data integration approach Bond *et al.* [8] combine open-licensed WordNet, Wiktionary[3] and Unicode Common Locale Data Repository CLDR[4] to link and extend the Open Multilingual WordNet (OMW). For the automatic methods, Montazery *et al.* [17] build a Persian WordNet relying on bilingual dictionaries as well as Persian and English monolingual corpus. Lam *et al.* [15] build Arabic, Assamese, Dimasa, Karbi, and Vietnam WordNet using the vocabulary of other languages and Microsoft's online machine translation system. They use the direct translation, intermediate WordNets, intermediate WordNets,

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

[4] http://cldr.unicode.org

and a dictionary (IWND) to create the multilingual WordNet synsets. Tarouti *et al.* [23] improve the above-mentioned methods by calculating word similarity based on word vectors, which improves the accuracy of the automatically constructed Arabic WordNet. Arcan et al. [7] use a multilingual parallel corpus for sense disambiguation, extend four European languages WordNets. Khodak *et al.* [14] automatically build WordNets through an unsupervised method based on word embedding and word induction, which utilizes machine translation and large-scale unlabeled monolingual corpus. Ercan and Haziyev [9] propose a graph clustering method in sense detection and apply it to multilingual WordNet construction.

In Uyghur language processing, there are some works about semantic resource construction. Aierken *et al.* [6] by using CUDD to construct a Uyghur language semantic lexicon based on PWN. The semantic lexicon contains 1300 pairs of synonyms and 1059 pairs of antonyms. Yilahun *et al.* [27] investigate Uyghur ontology construction methods and attempt to construct Uyghur ontology in computer science and mathematics. Qiu *et al.* [21] review construction of Uyghur knowledge graph. They also design and implement Uyghur-Chinese-English online dictionary based on knowledge graph[19] then apply the K-means method to semantic search [20]. Maimaiti *et al.* [16] construct named entity corpus from parallel corpus. Abiderexiti *et al.* [1,2] propose named entity relation annotation specification and construct a corresponding corpus. However, there are no reports about publicly available English-Uyghur WordNet datasets. As a result, it is hard to evaluate the accuracy and precision of automatically constructed English-Uyghur WordNet.

3 Dataset

3.1 Data Processing

CUDD mainly uses dictionary form of a word (lemma) as the description object and explains each sense of the word. Every sense has an explanation, and most senses have example sentences as shown in Figure 1. However, the granularity of each sense in CUDD is rougher than PWN. In general, a Uyghur word-sense pair corresponds to one or more than one word-sense pairs in PWN 3.0. To construct high-quality Uyghur-English WordNet, we preprocess CUDD and extract all the one sense Uyghur nouns from CUDD, whose total number is 10,973.

To get strong initial mapping relations between PWN 3.0 and CUDD, we need Uyghur-English and English-Uyghur dictionaries. But there are no off-the-shelf Uyghur-English or English-Uyghur dictionaries. We crawl Uyghur-English and English-Uyghur online dictionary from two websites (http://dict.izda.com, https://panlex.org). We have used 1s time interval for every word to not affect their server and to avoid our IP bing blocked. However, in some dictionary entries, there are irregular patterns and errors. Some Uyghur words are written in Latin and even Cyrillic script. Therefor, we first, eliminate these irregular patterns and errors manually. Then we convert all Latin and Cyrillic scripts into Uyghur scripts based on the Arabic alphabet which is the current writing

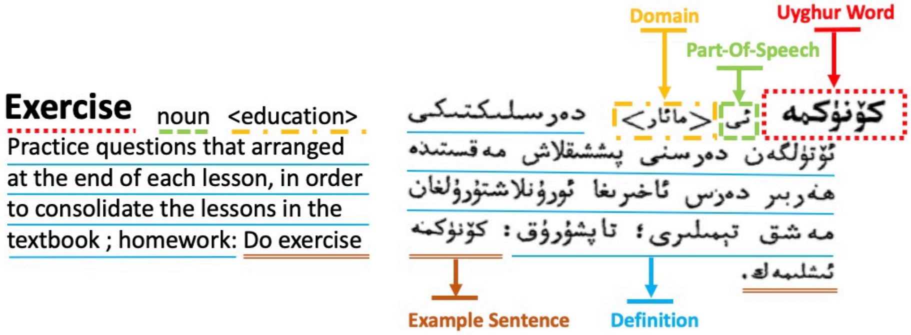
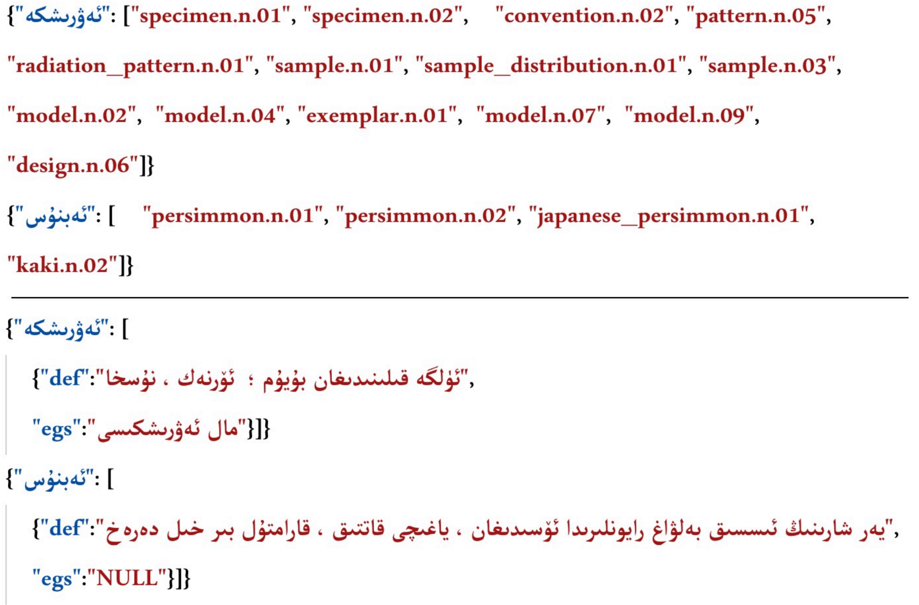

Exercise
noun
<education>
Practice questions that arranged at the end of each lesson, in order to consolidate the lessons in the textbook ; homework: Do exercise
Domain
Uyghur Word
Part-Of-Speech
كۆنۈكمە
Example Sentence
Definition

Fig. 1. An example from Contemporary Uyghur Detailed Dictionary (CUDD). The word in the red box with round dot is an Uyghur word, the character in the green box with dash line denote Part-Of-Speech of the word, the words in the orange box with dash dot line denote the domain of the word, the sentence marked with the blue line is a definition of the word, The sentence marked with the brown double line is an example sentence of the word.


```
{"ئەۋرىشكە": ["specimen.n.01", "specimen.n.02",    "convention.n.02", "pattern.n.05",
"radiation_pattern.n.01", "sample.n.01", "sample_distribution.n.01", "sample.n.03",
"model.n.02",  "model.n.04", "exemplar.n.01",  "model.n.07",  "model.n.09",
"design.n.06"]}
{"ئەبنۇس": [    "persimmon.n.01", "persimmon.n.02", "japanese_persimmon.n.01",
"kaki.n.02"]}
```

```
{"ئەۋرىشكە": [
  {"def":"ئۆلگە قىلىنىدىغان بۇيۇم ؛ ئۆرنەك ، نۇسخا",
  "egs":"مال ئەۋرىشكىسى"}]}
{"ئەبنۇس": [
  {"def":"يەر شارىنىڭ ئىسسىق بەلۋاغ رايونلىرىدا ئۆسىدىغان ، ياغىچى قاتتىق ، قارامتۇل بىر خىل دەرەخ",
  "egs":"NULL"}]}
```

Fig. 2. The JSON file format of the dataset: Arabic script in blue is an Uyghur word written in the current Uyghur writing system which is based on the Arabic alphabet. The upper part of the Figure indicates the relation between words in Contemporary Uyghur Detailed Dictionary (CUDD) and synsets in PWN 3.0 . The lower part of the Figure indicates the definition (denoted by "def") and example sentences (denoted by "egs") of the corresponding Uyghur word.

system in the Xinjiang Uyghur Autonomous Region. We also use some offline dictionaries which are typed manually. Finally, we combine these dictionaries to get the largest English-Uyghur and Uyghur-English dictionaries by exploiting many possible online and offline resources. The vocabulary sizes are 222,842 and 230,287 , respectively.

3.2 Building the Dataset

We denote Uyghur nouns as $W = \{w_1, w_2, \cdots, w_n\}$, where w_i to represent a word. T represents the English translation set of W, i.e. $T = \{t_1, t_2, \cdots, t_m\}$, Uyghur-English dictionary mapping function f_{ue} is:

$$f_{ue}(w_i) = T_i \tag{1}$$

where $T_i \subset T$, $T_i = \{t_{i1}, t_{i2}, \cdots, t_{io}\}$, $T_i \neq \emptyset$.

We denote $S = \{s_1, s_2, \cdots, s_p\}$ to represent the synsets in PWN. The relationship between English words and its synset is based on the following equation (2) :

$$P_{syn}(T_i) = S_i \tag{2}$$

where $S_i \subset S$, $S_i = \{s_{i1}, s_{i2}, \cdots, s_{io}\}$.

For every word $w_i \in W$, we get the corresponding English synsets S_i by equation (1) and (2). In this way, we build first English-Uyghur WordNet raw mapping data, and the synsets are 73,491 . We split the data into three parts for training, developing, and testing. We annotate the 20,702 English synsets manually according to the definitions and example sentences of both Uyghur and English. It is hard to find someone who is proficient in two languages and also have linguistics and computer science background. So one person who satisfies above condition annotate the synsets two times and then rechecked it. Because one person annotates the synsets, so we do not calculate the annotation agreements between two times annotations. Finally we get 6,642 English synsets and corresponding 1,750 Uyghur by checking three times. The statistics of the training set, development set, and test set is shown table 1.

Table 1. Evaluation dataset for automatic English-Uyghur WordNet construction

	# of training set	**# of development set**	**# of test set**
English	52,789	5,493	1,149
Uyghur	4,376	1,450	300

We make our automatically and manually labeled dataset publicly available (https://github.com/kaharjan/uywordnet). The Uyghur words include the meaning of words and examples, the format of the file is JSON which is shown by Figure 2. Each of the training, development, and test set contains two files. One is CUDD and PWN 3.0 mapping relations shown in the upper part of Figure 2. Another is the definition and examples of Uyghur word in CUDD shown in the lower part of Figure 2.

4 Mapping Method

4.1 Uyghur Word Representation

To map corresponding English synsets to Uyghur words, We take advantage of the idea of word embeddings, which allows words with similar meaning to have a similar word representation. We have used the monolingual corpus, including more than three million sentences, which are the largest corpus in Uyghur. However, for the low resource language and agglutinative language like Uyghur, beside the corpus size, getting reasonable word representation also faces challenges. In Uyghur, one word could have several suffixes to form phrases, even sentences. For example, " بېيجىڭدىكىلەرنىڭكىمۇ؟ " (Does it belongs to people who are from Beijing?). So getting word representation in forms of the dictionary would be difficult. Fortunately, morpheme-enhanced continuous bag-of-words model (mCBOW) [3] cope with this problem. The idea of mCBOW is to treat morphemes rather than words as the basic unit of representation learning. To use mCBOW, we need to stem or lemmatize Uyghur words. Although there are some works about Uyghur word morphological segmentation [5][18] and annotated corpus [4], no off-the-shelf stemming, lemmatization tool or large scale annotated corpus for practical use. So we have applied two existing solutions. First is probabilistic generative models that use sparse priors inspired by the Minimum Description Length (MDL) principle [24][11]. The second is the subword model based on the byte pair encoding compression algorithm[22]. We conduct several experiments to compare and tune the models, then use the best solution to train the mCBOW model to get optimal word representation. The comparing experiment results will discuss in section 5.

4.2 Synset Mapping based on Word Embeddings

After getting optimal results for word representation, we use the training data described in section 3 to get the corresponding lemmas in English. More formally, We denote P_{lemm} to represent the mapping function from synsets to lemmas in equation (3) :

$$P_{lemm}(S_i) = L_i \tag{3}$$

where L_i represents a subset of the lemma set L, $L_i = \{l_{i1}, l_{i2}, \cdots, l_{io}\}$. L_i is a true subset of L $L_i \subset L$.

We derive the Uyghur word set W_i from the English-Uyghur dictionary mapping function f_{eu} according to equation (4):

$$f_{eu}(L_i) = W_i \tag{4}$$

where $W_i = \left\{W^{'}_{i1}, W^{'}_{i2}, \cdots, W^{'}_{io}\right\}$,$W^{'}_{ij} \subset W$.

The word embeddings of Uyghur word $\boldsymbol{v}_{w_i}$ is calculated by equation (5):

$$\boldsymbol{v}_{w_i} = \boldsymbol{u}^{'}_{w_i} + \frac{1}{m}\sum_{k=1}^{m} \boldsymbol{u}_k \tag{5}$$

where $\boldsymbol{u}^{'}_{w_i}$ is representation of surface form of the word w_i, $\boldsymbol{u}_k$ is representation of the kth unit of the word w_i. In other words, w_i consists of m units.

We calculate the cosine similarity between each Uyghur word in the set $W_i^{'}$ and the word vector corresponding to w_i in equation (6):

$$Sim(w_i, W_{ij}^{'}) = \frac{1}{|W_{ij}^{'}|} \sum_{w_{ijk}^{'} \in W_{ij}^{'}} \cos(\boldsymbol{v}_{w_{ijk}^{'}}, \boldsymbol{v}_{w_i}) \tag{6}$$

where $\boldsymbol{v}_{w_i}$ and $\boldsymbol{v}_{w_{ijk}^{'}}$ are the vectors of the Uyghur words w_i and $w_{ijk}^{'}$. $|W_{ij}^{'}|$ the number of the words in the Uyghur word set $W_{ij}^{'}$ corresponding to Uyghur word w_i. Finally, we get $S_i^{'} = \{s_{i1}, s_{i2}, \cdots, s_{ip}\}$, $0 \leq p \leq o$ as the mapping result for w_i . $S_i^{'}$ is English synsets corresponding to the Uyghur words in $match(w_i)$ obtained by equation (7). The lower and upper bound of synset number p is between 0 and o.

$$match(w_i) = \underset{W_{ij}^{'} \in W_i}{\arg\max}\, Sim(w_i, W_{ij}^{'}) \tag{7}$$

5 Experiments

5.1 Word Representation

As discussed above section, first we evaluate stemming for mCBOW. We separately train the models based on Morfessor 2.0 [24] and FlatCat[11] on the same corpus. We use two different corpora for training. The one is Uyghur text in parallel Chinese-Uyghur Corpus[5], the Uyghur is the translation of Chinese text that most of them are news and laws. The other one is the part of our Uyghur monolingual muli-domain corpus. To compare fairly, we set the size of the monolingual corpus equal to the previous corpus. For semi-supervision, we use the annotated corpus in [4]. We conduct several experiments about the different parameters of Morfessor 2.0 and FlatCat. In the experiments, we use test data in[5] and the standard precision, recall, and F1 score as the primary performance indicators to evaluate stemming. Note that, we do not compare the models of Morfessor 2.0 and FlatCat with the models of BPE. Because BPE controls the granularity of the segmented unit by operation number, so we directly use this model in our word representation. The table 2 is comparison of the various models of Morfessor 2.0 and FlatCat.

The table 2 shows that although the FlatCat model based on semi supervision (Mono_fc_semi) has achieved the highest precision, the recall is much lower than other models of Morfessor 2.0. Initially, the models based on Morfessor 2.0 are chosen. Then we compare models based on Morfessor 2.0. In the Morfessor 2.0 models, the semi-supervised methods are better than the unsupervised method (CLDC_mf_unsup). There is not much significant difference between the model

[5] http://www.chineseldc.org

Table 2. The Comparison between various stemming models

Models	**P**	**R**	**F1**
CLDC_mf_unsup	0.827	0.436	0.571
CLDC_mf_semi	0.802	0.729	0.764
MONO_mf_semi	0.819	0.733	**0.774**
MONO_fc_semi	0.999	0.207	0.343

trained on the monolingual corpus (Mono_mf_semi) and the model based on translation corpus (Mono_mf_semi). So, finally, we choose the two models, *Mono_mf_semi* and *Mono_mf_semi* , to train mCBOW.

We evaluate the performance of word embeddings on the *uyWordSim-196* dataset which is a subset *uyWordSim-353* [3]. The *uyWordSim-353* is Uyghur translation of popular English *WordSim-353* dataset. In the translation of the *WordSim-353*, some of the words in English corresponding to two or more Uyghur words which would become OOV words in training data without multi-word expression identification. So [3] filtered 196 pairs that corresponding to one Uyghur words in *uyWordSim-353* to form *uyWordSim-196*. They mainly test the performance of the word representation on the *uyWordSim-196*. So we also choose this subset as our benchmark. The result of Spearman (ρ) correlation to human judgment is shown in the table 3 :

Table 3. Spearman (ρ) correlation results for word similarity on the *uyWordSim-196* dataset

Models	***uyWordSim-196***
mCBOW_CLDC_mf_semi	50.99
mCBOW_Mono_mf_semi	51.73
mCBOW_BPE_32K	58.26
mCBOW_BPE_100K	**63.13**
mcBOW_BPE_150K	62.15

As the table 3 is shown, we segment the text applying two categories of models which are based on Morfessor 2.0 and BPE. We set the operation number of BPE 32K, 100K, and 150K. The result shows that the mCBOW based model is depended on segmentation, the highest score on *uyWordSim-196* is 63.13. The BPE based model is better than the Morfessor 2.0 based model. When the operation number is 32K or 150K, word segmentation granularity is too coarse or too fine, the performance suffers from low precision. When the operation number is 100K, the model performs best. So we choose BPE based mCBOW for the OOV problem in Uyghur.

5.2 Synset Mapping based on Word Embdeddings

We evaluate our Synset Mapping based on Word Embeddings (SMWE) approach on our dataset, which will be publicly available for other's comparative research. We tune the similarity threshold by using the development set. On the test set our approach 59.24 precision, 73.06 recall, and 65.03 F1-score. It can be seen that the SMWE make it possible to construct Uyghur WordNet automatically using only monolingual corpus and bilingual dictionaries.

5.3 Case Study

To demonstrate the effectiveness of the synset mapping method, we provide instances in our test dataset for example shown by Table 4. The Uyghur word "پەرمانچى" and English synset "herald.n.01", "harbinger.n.01" map each other in raw mapping. SMWE able to map correctly. In the second example, Uyghur word " گۈلزار " is mapped to wrong synsets like "modling.n.03", "border.n.05", but SMWE still maps correctly. However, in the third example Uyghur word " تۈزۈت " is mapped to "embellishment.n.01", "embroidery.n.02", which is correct. But according to both definition of English and Uyghur, it is missed the correct one "civility.n.01" . So our method still some rooms for improvement.

6 Conclusion

The main contribution of this paper is that we construct an English-Uyghur WordNet dataset for automatically evaluation. In the first time make this dataset available publicly (https://github.com/kaharjan/uywordnet). In this dataset, 73,491 English synsets in PWN 3.0 are mapped to Uyghur in Contemporary Uyghur Detailed Dictionary (CUDD), 20,702 are manually annotated to get 6,642 English synsets. The annotated data are divided into developing and test for the evaluation of English-Uyghur WordNet synset mapping methods. In this dataset, every Uyghur word has definitions or examples in the Uyghur language context, which make this dataset more valuable for downstream application. During these works, we build the largest Uyghur-English and English-Uyghur dictionary using all available online and offline resources, which put a considerable amount of work. We get better word representation of Uyghur using the largest monolingual corpus. This word representation alleviates the OOV problem based on previous work. Also, we propose an English-Uyghur synset mapping based on word embeddings (SMWE) method. The experimental results on the dataset are promising. Because there is no machine translation (MT) between English and Uyghur, we do not exploit the definition of Uyghur words and examples in our SMWE. In the future, we will further investigate how to expand our dataset and improve the performance of our algorithm.

Acknowledgments

This work is supported by National Natural Science Foundation of China (NSFC) grant 61532001.

Table 4. Examples from test dataset for synset mapping method based on word embbeddings (SMWE)

Uyghur Word	Definition	Raw Mapping	SMWE	Gold Standard
پەرمانچى	پەرماننى باشقىلارغا يەتكۈزگۈچى ؛ جاكارلىغۇچى	herald.n.01, harbinger.n.01	herald.n.01	herald.n.01
گۈلزار	ئۆستۈرۈلگەن ، گىياھلار - گۈل قاپلانغان بىلەن جاي ، گۈللۈك ، چىمەنزار	flowerbed.n.01 , flower_ garden.n.01 , boundary_ line.n.01, margin.n.01 , edge.n.01 , molding.n.03 , border.n.05	flowerbed.n.01 , flower_ garden.n.01	flowerbed.n.01, flower_ garden.n.01
تۈزۈت	ئارتۇقچە تارتىنچاقلىق ، بىلەن تەكەللۇپ مۇئامىلە قىلغان	politeness.n.01, politeness.n.02, civility.n.01, cold.n.01, coldness.n.03, cold.n.03, coldness.n.02, frigidity.n.01, chill.n.01, frisson.n.01, chill.n.03, chill.n.04	politeness.n.01, politeness.n.02	politeness.n.01, politeness.n.02, civility.n.01

References

1. Abiderexiti, K., Maimaiti, M., Yibulayin, T., Wumaier, A.: Annotation Schemes for Constructing Uyghur Named Entity Relation Corpus. In: The 2016 International Conference on Asian Language Processing (IALP 2016). pp. 103–107 (2016)
2. Abiderexiti, K., Maimaiti, M., Yibulayin, T., Wumaier, A.: Construction of Uyghur Named Entity Relation Corpus. International Journal of Asian Language Processing **27**(2), 155–172 (2017)
3. Abudukelimu, H., Liu, Y., Chen, X., Sun, M., Abulizi, A.: Learning distributed representations of uyghur words and morphemes. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **9427**, 202–211 (2015)
4. Abudukelimu, H., Sun, M., Liu, Y., Abulizi, A.: THUUyMorph:An Uyghur Morpheme Segmentation Corpus. Journal of Chinese Information Processing **32**(02), 81–86 (2018) (In Chinese)

5. Abudukelimu, H., Cheng, Y., Liu, Y., Sun, M.: Uyghur morphological segmentation with bidirectional GRU neural networks. J Tsinghua Univ(Sci & Technol) **57**(1), 1–5 (2017) (In Chinese)
6. Aierken, R., Xiao, L., Tohti, A., Jiang, Z.M.: Constructing a Uyghur language semantic lexicon based on WordNet. In: 2014 Science and Information Conference. pp. 182–186 (2014)
7. Arcan, M., McCrae, J.P., Buitelaar, P.: Expanding wordnets to new languages with multilingual sense disambiguation. In: Proceedings of COLING 2016: Technical Papers. pp. 97–108 (2016)
8. Bond, F., Foster, R.: Linking and extending an open multilingual Wordnet. In: Proceedings of ACL 2013. pp. 1352–1362 (2013)
9. Ercan, G., Haziyev, F.: Synset expansion on translation graph for automatic wordnet construction. Information Processing and Management **56**(1), 130–150 (2019)
10. Fellbaum, C.: WordNet. In: Theory and Applications of Ontology: Computer Applications, pp. 231–243. Springer (2010)
11. Grönroos, S.A., Virpioja, S., Smit, P., Kurimo, M.: Morfessor FlatCat: An HMM-Based Method for Unsupervised and Semi-Supervised Learning of Morphology. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics. pp. 1177–1185 (2014)
12. Huang, C.R., Chang, R.Y., Lee, S.B.: Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004). pp. 26–28. Lisbon, Portugal (2004)
13. Huang, C., Hsieh, S., Hong, J., Chen, Y., Su, I., Chen, Y., Huang, S.: Chinese wordnet : design, implementation, and application of an infrastructure for cross-lingual knowledge processing. Journal of Chinese Information Processing **24**(02), 14–23 (2010) (In Chinese)
14. Khodak, M., Risteski, A., Fellbaum, C., Arora, S.: Automated WordNet Construction Using Word Embeddings. In: Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications. pp. 12–23 (2017)
15. Lam, K.N., Al Tarouti, F., Kalita, J.: Automatically constructing Wordnet Synsets. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 106–111 (2014)
16. Maimaiti, M., Wumaier, A., Abiderexiti, K., Wang, L., Wu, H., Yibulayin, T.: Construction of Uyghur named entity corpus. In: Yang, E., Sun, L. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (2018)
17. Montazery, M., Faili, H.: Automatic Persian WordNet Construction. In: Coling 2010: Poster. pp. 846–850. Beijing,China (2010)
18. Osman, T., YANG, Y., Tursun, E., CHENG, L.: Collaborative Analysis of Uyghur Morphology Based on Character Level. Acta Scientiarum Naturalium Universitatis Pekinensis **55**(01), 47–54 (2019) (In Chinese)
19. Qiu, L., Yang, H., Zhou, R.: The Design and Implementation of Chinese-Uighur-English Online Dictionary Based on Knowledge Graph. In: 22017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC). pp. 883–886 (2017)
20. Qiu, L., Yang, N., Maolimamuti, M.: Chinese-Uyghur-English Semantic Search Based on the Knowledge Graphs. In: 2017 IEEE International Conference on Com-

putational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC). pp. 879–882 (2017).

21. Qiu, L., Zhang, H.: Review of Development and Construction of Uyghur Knowledge Graph. 22017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC) pp. 894–897 (2017).
22. Sennrich, R., Haddow, B., Birch, A.: Neural Machine Translation of Rare Words with Subword Units. In: Proceedings ofthe 54th Annual Meeting ofthe Association for Computational Linguistics. pp. 1715–1725. Berlin,Germany (2016)
23. Tarouti, F.A., Kalita, J.: Enhancing Automatic Wordnet Construction Using Word Embeddings. In: Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP. pp. 30–34. Association for Computational Linguistics (2016)
24. Virpioja, S., Smit, P., Grönroos, S.A., Kurimo, M.: Morfessor 2.0 : Python Implementation and Extensions for Morfessor Baseline. Tech. rep., Aalto University, School of Electrical Engineering, Department of Signal Processing and Acoustic (2013)
25. Wang, S., Bond, F.: Building the Chinese Open Wordnet (COW): Starting from Core Synsets. In: International Joint Conference on Natural Language Processing. pp. 10–18. Asian Federation of Natural Language Processing, Nagoya, Japan (2013)
26. Xu, R., Gao, Z., Pan, Y., Qu, Y., Huang, Z.: An Integrated Approach for Automatic Construction of Bilingual Chinese-English WordNet. In: Domingue, J., Anutariya, C. (eds.) Asian Semantic Web Conference. vol. 5367, pp. 302–314. Springer Berlin / Heidelberg (2008).
27. Yilahun, H., Imam, S., Hamdulla, A.: A Survey on Uyghur Ontology. International Journal of Database Theory and Application **8**(4), 157–168 (2015)