# Construction of Uyghur named entity corpus

**Maihemuti Maimaiti[1,2], Aishan Wumaier[1,2], Kahaerjiang Abiderexiti[1,2,]**
**Wanglulu[1,2], Wuhao[1,2], Tuergen Yibulayin[1,2]**

[1] School of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang 830046, China
[2] Laboratory of Multi-Language Information Technology, Urumqi, Xinjiang 830046, China
mahmutjan@xju.edu.cn, hasan1479@xju.edu.cn, kaharjan@xju.edu.cn,wanglulu@stu.xju.edu.cn,
840315259@qq.com, turgun@xju.edu.cn

## Abstract

The research of named entity recognition plays an important role in natural language processing (NLP) and can improve the performance of high-level NLP tasks, such as Machine Translation, Question Answering System, syntactic analysis and so on. Uyghur is a morphologically rich language with lack of manually created language resources. In this paper, we present the construction process of Uyghur named entity annotated corpus. This process composes of three steps. In the first step Chinese named entity recognition is used to select sentences from sentence aligned corpus and to extract Chinese named entities. In the second step Chinese-Uyghur named entity dictionary is automatically constructed using Chinese-Uyghur machine translation system, for the automatic pre-annotation of Uyghur named entities, and all annotations are corrected manually using an annotation tool. In the final step, corpus annotation quality is improved by using multiple strategies. As a result, four different sentence-level annotated corpora, person name annotated corpus, location name annotated corpus, organization name annotated corpus and personal, location, organization names annotated corpus, are constructed separately. To our knowledge, this is the first large-scale Uyghur named entity annotated corpus(UNEC) which is very valuable for the further researches.

**Keywords:** Uyghur, named entity, corpus

## 1. Introduction

With the development of the Internet, more and more text data appear. Information extraction has become an important direction in the field of natural language processing. Among them, named entity recognition (Coates-Stephens S. 1992; Thielen C. 1999) is a hot spot of information extraction proposed by MUC-6 conference (Sundheim B M. 1996), which is an important part of natural language processing. There are a large number of entities in the text, such as person names, location names and organization names. With the increase of text data, there are constantly appearing new named entities, and some of the named entities may be eliminated. It is almost impractical to construct a dictionary containing all named entities. Therefore, automatic recognition of named entities is an important task, other entities are easily treated as unknown words in the processing of natural language processing. Thus, affecting the performance of machine translation, knowledge map construction, Question Answering System, syntax analysis and other application areas. There are a large number of named entity annotation corpora in languages such as English (Sang E F T K et al. 2003) and Chinese (Walker C et al. 2006) at present. The named entity recognition technologies in many languages are relatively mature (Polifroni J et al. 2010; Savary A et al. 2010; Desmet B et al. 2013). But resource-deficient languages, like Uyghur, so far, no publicly available named entity corpus has yet to appear. On the one hand, it limits the research of Uyghur named entity recognition, on the other hand, it has some influence on the development of Uyghur information extraction technology.

Therefore, this paper firstly collected a large number of bilingual corpora in the field of news. It explores how to quickly establish a named entity corpus, for a resource-deficient language, by using cross-language named entity recognition technology. Firstly, named entities are automatically labeled for resource-rich languages. Secondly, sentence pairs with named entities are selected and pre-labeled for resource-deficient language sentences using bilingual named entity dictionaries. Finally, corrections and supplements were made manually, and annotation memory technology was used to further improve the efficiency and quality of annotation. Thus, Uyghur language location name annotation corpus, organization name annotation corpus, person name annotation corpus and Uyghur named entity annotation corpus are constructed respectively.

## 2. Related work

Compared with Chinese and English, the construction of the Uyghur Named entity corpus is very backward. In the study of Uyghur named entity and the construction of Uyghur annotated corpus, researchers have already made some contributions. Here are some important jobs, according to the features of Uyghur person name, this work built Uyghur person name annotated corpus that contained 5258 sentences (Rozi A et al. 2013). In the integrated recognition task which includes person name, organization name and location name, they constructed the a comprehensive named entity corpus that consists of 11257 annotated sentences (Tashpolat N et al. 2017). In addition, researchers constructed Uyghur music entity corpus which contains 2400 sentences in the research of Uyghur music entity recognition (Adila Ahmat et al.2017). The above are all Uyghur named entity corpus that have been reported. At present, the number of Uyghur entity annotated corpora is very small, and most of researchers used the data from the internet to research by rule-matching (Arkin M et al. 2013; Maihefureti et al. 2014; Mahmoud A et al. 2017). However, the research of the named entity recognition by machine learning (Yang Y et al. 2011; Jiazheng L I et al. 2011; Abiderexiti K et al. 2017) can't be separated from the standard data resource. So it is necessary to build standard data resource of named entity.

# 3. Creating the data resource

The construction of a corpus requires not only a large amount of data resource, but also expensive manpower. The details of construction are as follow.

## 3.1 Data Source of corpus

Based on the Uyghur-Chinese parallel corpus of the 13th China Workshop on Machine Translation (CWMT2017), we also use the Uyghur-Chinese parallel corpus provided by the Laboratory of Xinjiang Multilingual Information Technology for manual annotation

## 3.2 Annotation specification

In the processing of manual annotation, for all corpora, we use three kinds of tags, Person (PN, personal name), Location (LN, place name) and Organization (ON, organization name). Due to one sentence may contain two or more tokens, this paper adopts annotation specification named BIO (begin-in-out) proposed by (Ramshaw L A et al. 1995). The tagging sets contains 7 kinds of labels. Table 1 shows this tag sets:

| label | Meaning |
|---|---|
| O | Non entity words |
| B-PER | The first word of person name or person name of a single word |
| I- PER | Not the first word in the person name |
| B-LOC | The first word of location name or location name of a single word |
| I-LOC | Not the first word in the location name |
| B-ORG | The first word of organization name or organization name of a single word |
| I-OR | Not the first word in the organization name |

Table 1: Tag sets of named entity in Uyghur

This is an example to describe an original sentence and tagged sentence , as Table 2 shown:

| | |
|---|---|
| **Original sentence** | Shi Jinping Bëyjingda Birleshken Döletler Teshkilatining bash katipi Ban Kimon bilen körüshti .<br>(Translation : Xi Jinping met with UN Secretary-General Ban Ki-moon in Beijing.) |
| **Tagged sentence** | 【Shi Jinping PN】【Bëyjingda LN】【 Birleshken Döletler Teshkilatining ON】 bash katipi 【Ban Kimon PN】 bilen körüshti . |
| **After tagging** | Shi/B-PER Jinping/I-PER Bëyjingda/B-LOC Birleshken/B-ORG Döletler/I-ORG Teshkilatining/I-ORG bash/O katipi/O Ban/B-PER Kimon/I-PER bilen/O körüshti/O ./O |

Table 2: An examle of tagged sentence

## 3.3 Annotation method and processing

In order to reduce the work of manual annotation, we used a method of man-machine combination to improve the speed of tagging and guarantee the quality of corpus at the same time. Processing of construction incorporated three steps: preprocessing, tagging, post-processing. The whole processing is showed as Figure 1.

### 3.3.1 Pre-processing

**Bilingual sentence deduplication processing**: Due to the bilingual sentence alignment corpora is very large, there may be repeated sentences. To avoid re-labeling the same sentence, it is necessary to remove the repeated sentences. There are two steps, the first step is to remove all the punctuation, only to reserve Chinese character, then removing the repeated sequence of Chinese characters. The second step is to do the same operation in Uyghur.

**Chinese named entity annotation**: After the processing of bilingual sentence deduplication, the NLPIR[1] system was used for construction of the location name annotated corpus, the other named entity corpus used Pyltp[2] system which from Harbin Institute of Technology. The functions of these two systems incorporate Chinese word segmentation and Named entity annotation.

**Entity extraction and sentence filtering**: In order to reduce the manual checking time and improve the speed of tagging, we filter out sentences which not incorporate person name, location name and organization name. Then, corpus only contained the Uyghur sentences which prepare to tag and the corresponding Chinese named entity list which include entity labels.

### 3.3.2 Name entity tagging

**Chinese and Uyghur name entity dictionary automatic construction:** First, we redo the Chinese named entity list and construct the entity dictionary in descending order by the occurrence frequency. Our group's Chinese-Uyghur machine translation system had been used to translate this dictionary and generated a Bilingual entity dictionary. There are some problem in the dictionary, such as incorrect translation, empty translation result and translation result include a variety of additional components. Therefore we manually review the translation result and correct the entity dictionary.

**Named entity automatic annotation:** According to the established Chinese-Uyghur named entity dictionary, all Uyghur sentences had been tagged for the first time. In order to prevent the spread of errors, the Uyghur entities corresponding to the Chinese entity have been tagged based on Chinese entities list. Since named entity in Uyghur sentences may be variants that attached some additional component, we used a method of fuzzy matching.

**Named entity manual proofreading:** Through automatic tagging, part of Ughur sentences has been tagged well. There are still many problems in the result of automatic annotation, such as error of Chinese named entity recognition, the source sentence alignment problem (the content is not aligned, the translation longer or less than the original sentence, etc.) and variants of the named entity result in matching error. Due to these problems, all the automatic annotated Uyghur sentences were imported in manual annotation system and conducted manual

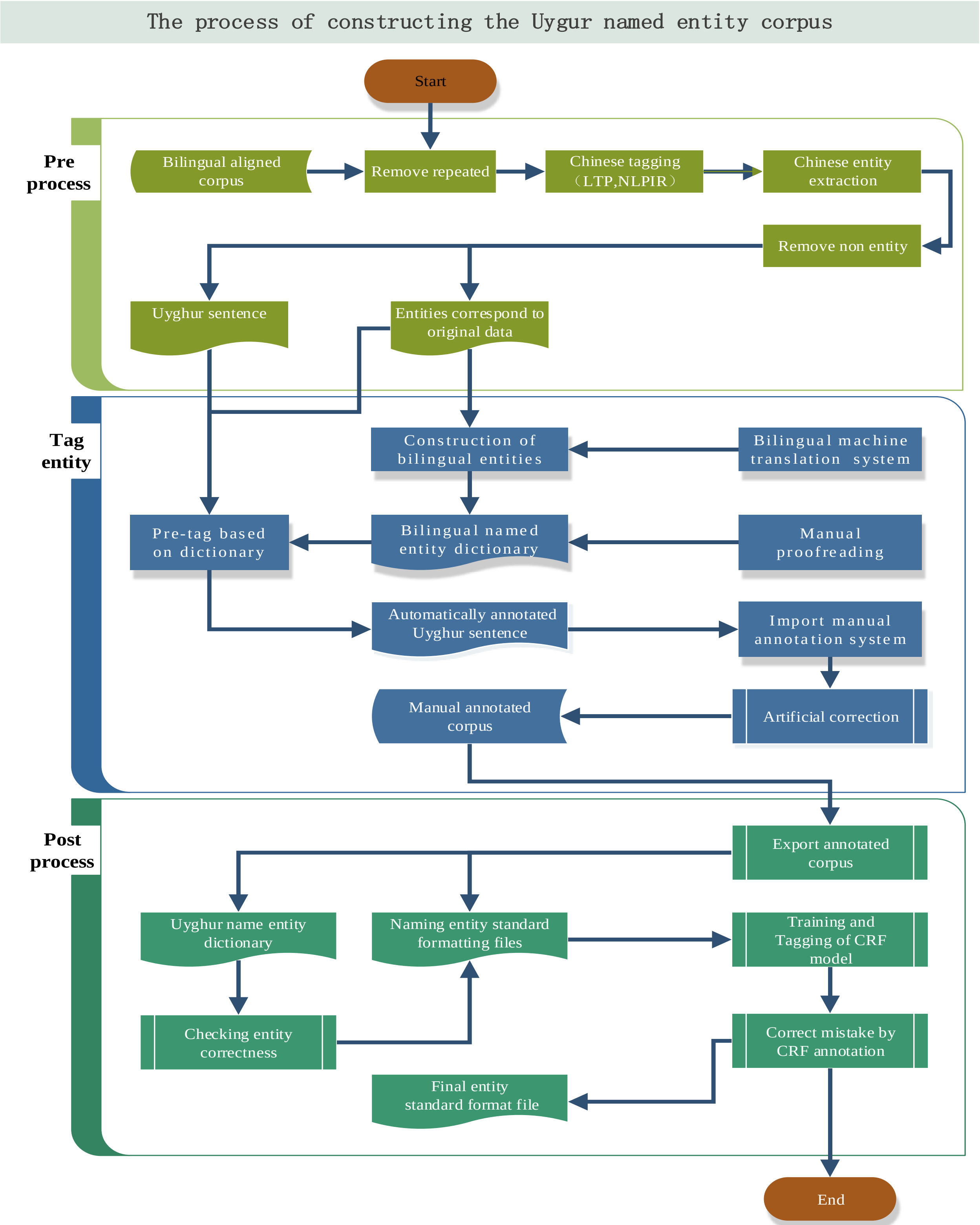
[1] https://github.com/NLPIR-team/NLPIR

[2] https://pypi.python.org/pypi/pyltp

proofreading and correction. The same named entity would be ignored to improve the speed of tagging and ensure the consistency of tagging.

### 3.3.3 Post-processing

Since multiple people tag at the same time, it is difficult to avoid mistakes just like annotation error, annotation inconsistency, leakage of annotation, etc. Based on the manual annotated corpus, we established Uyghur named entity dictionary and proofread each entity by using source sentences. After that, the CRF annotation machine was trained with all the language materials, and the corpus has been automatically tagged. Other possible error would be corrected by comparing with manual annotation and automatic annotation. Finally, we derive the named entity corpus in the format of the above tagged sentence.



The process of constructing the Uygur named entity corpus
Start
Pre process
Bilingual aligned corpus
Remove repeated
Chinese tagging （LTP,NLPIR）
Chinese entity extraction
Remove non entity
Uyghur sentence
Entities correspond to original data
Tag entity
Construction of bilingual entities
Bilingual machine translation system
Pre-tag based on dictionary
Bilingual named entity dictionary
Manual proofreading
Automatically annotated Uyghur sentence
Import manual annotation system
Manual annotated corpus
Artificial correction
Post process
Export annotated corpus
Uyghur name entity dictionary
Naming entity standard formatting files
Training and Tagging of CRF model
Checking entity correctness
Correct mistake by CRF annotation
Final entity standard format file
End

Figure 1: Flow chart of the construction of Uyghur named entity corpus.

## 4. Data analysis

The whole corpus resources contains location name annotated corpus, person name annotated corpus, organization name annotated corpus and a comprehensive named entity corpus include person name, location name, organization name.

### 4.1 Location name annotated corpus

Location name annotated corpus contains 13385 sentences with 20218 place names, and the number of location names appeared 41009 times. The statistics are shown in Figure 2. Most of location names are consist of one or two words, total occupying 93% of the location name number, and their number are 8506 and 10296 respectively, shown in Figure 3.

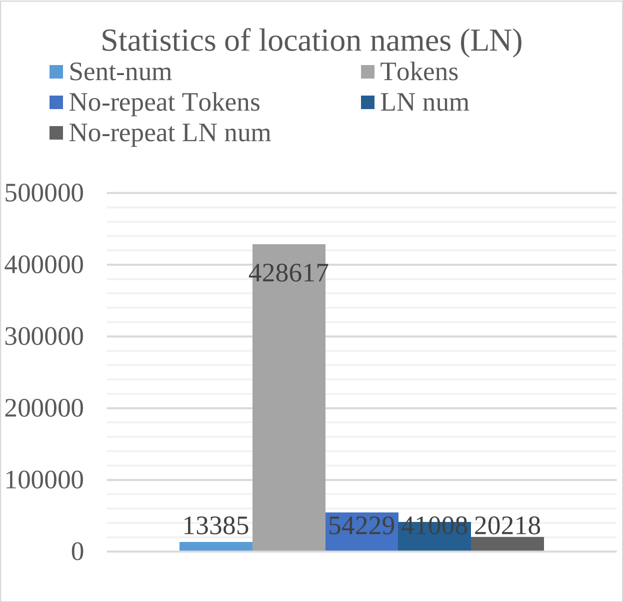


Figure 2: Location name annotated corpus statistics.

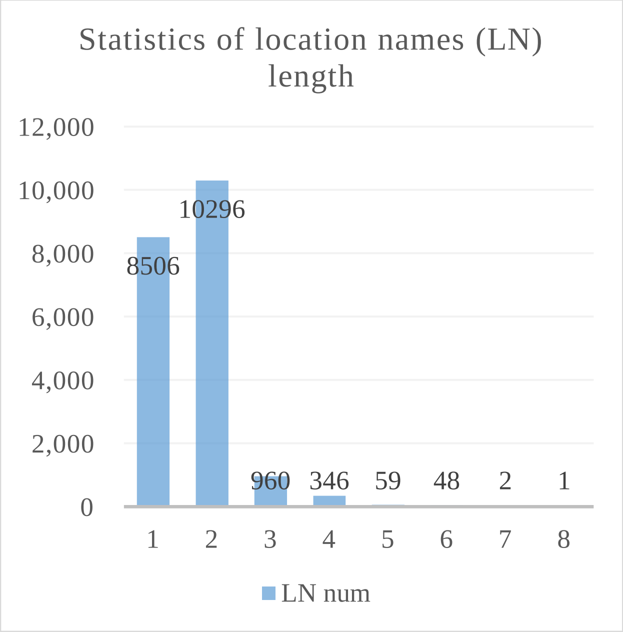


Figure 3: Location name length statistic.

### 4.2 Organization name annotated corpus

Organization name annotated corpus contains 11337 sentences with 9733 organization names, and the number of organization names appeared 13436 times. The specific statistics are shown in Figure 4.

The statistic of organization name annotated corpus found that the length of most organization name is between 2 and 11 words. Compared with the location name, the length of organization name is longer and the distribution is wider. The specific statistics are shown in Figure 5.

### 4.3 Person name annotated corpus

Person name annotated corpus contains 21078 sentences with 18066 person names， and the number of person names appeared 34598 times. As shown in the Figure 6:

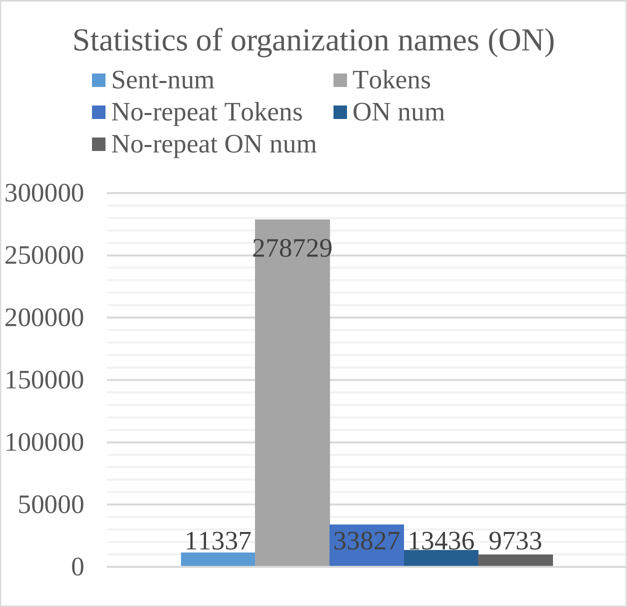


Figure 4: Organization name annotated corpus statistics.

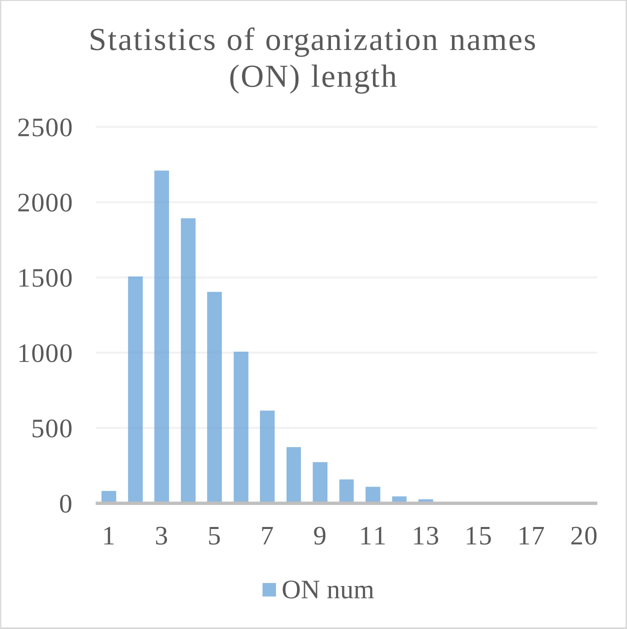


Figure 5: Organization name length statistic.

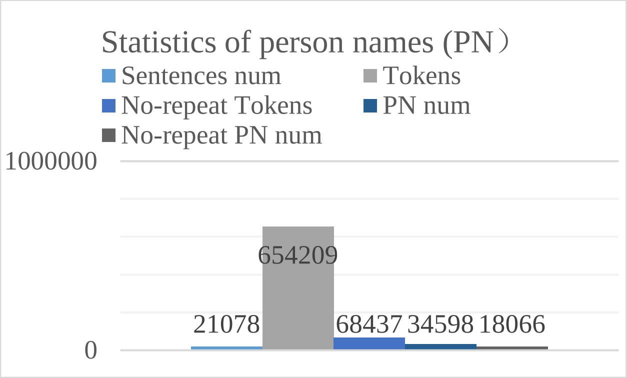


Figure 6: Person name annotated corpus statistic.

Most of person name are consist of one or two words, their proportion is high up to 99%. The Figure 7 show this situation.

## 4.4 Named entity annotated corpus

Uyghur named entity corpus which is comprehensive corpus contains 39027 sentences. The number of entities is 102360(include repeating entities). The person name appeared 28469 times, account for 27.8%, the location name appeared 42585 times, account for 41.6%, and organization names appeared 31306 times, account for 30.6%. The specific statistics are shown in Figure 8.

Figure 9 shows the length of different named entity in corpus. As can be seen from the statistics, the length of person name more than two words is rare; the length of location name is almost between 1 to 6 words, the number of location name was decrease with increase entity length; the length of organization name is mainly between 2 to 12 words, its length is relatively long.

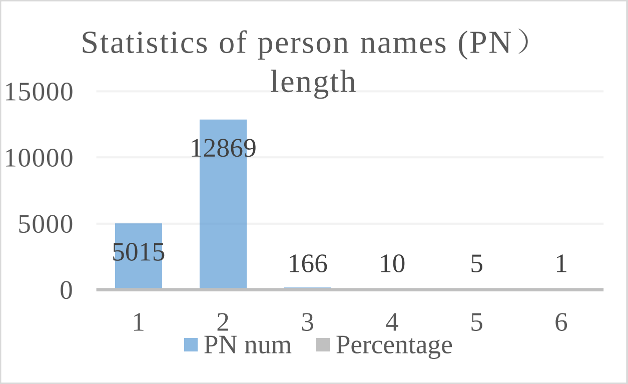
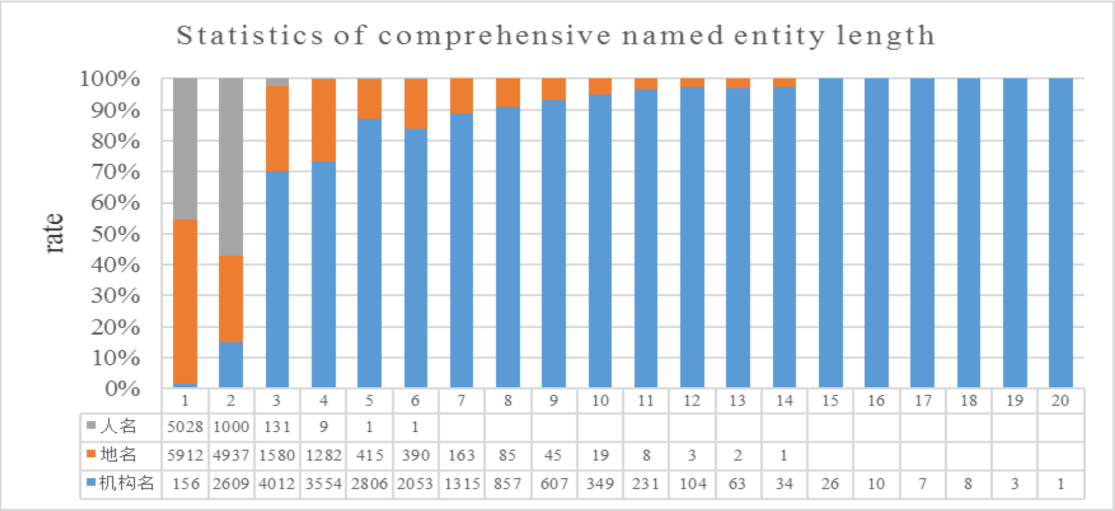


Figure 7: Person name length statistic.



| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ■人名 | 5028 | 1000 | 131 | 9 | 1 | 1 | | | | | | | | | | | | | | |
| ■地名 | 5912 | 4937 | 1580 | 1282 | 415 | 390 | 163 | 85 | 45 | 19 | 8 | 3 | 2 | 1 | | | | | | |
| ■机构名 | 156 | 2609 | 4012 | 3554 | 2806 | 2053 | 1315 | 857 | 607 | 349 | 231 | 104 | 63 | 34 | 26 | 10 | 7 | 8 | 3 | 1 |

Figure 8: Statistics of Uyghur named entity corpus.

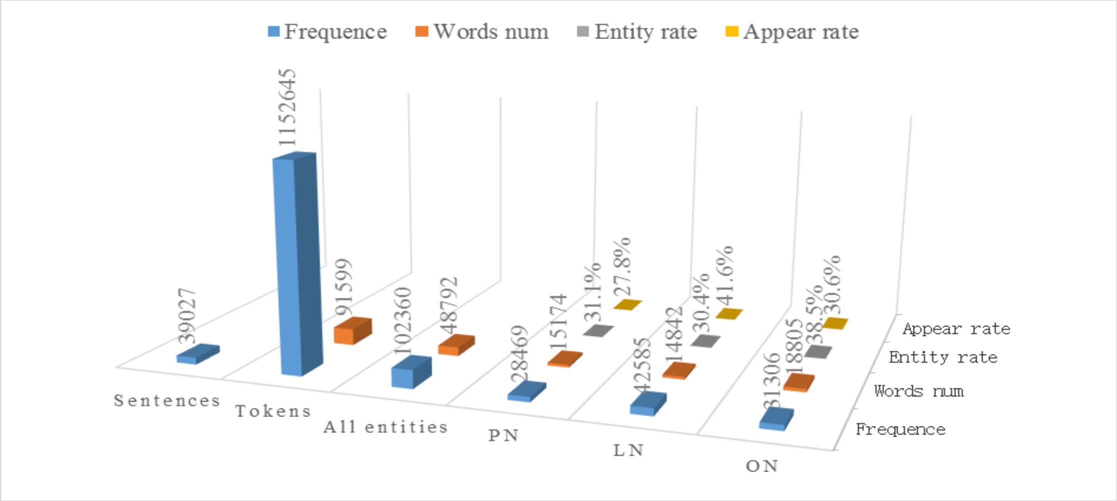


Figure 9: Uyghur named entity corpus length statistic.

## 5. Conclusions

Our paper used bilingual sentence alignment corpus and Chinese named entity recognition technology to establish person name annotated corpus, location name annotated corpus, organization name annotated corpus and named entity annotated corpus in Uyghur. During the construction of corpus, it is very useful to build bi-lingual entity dictionary, automatic annotation, annotation memory, error analysis and so on. By using a human-machine combination method which greatly reduces the cost of human resources and effectively guarantees the quality of the corpus.

With the completion of constructing corpus, our work has laid a solid foundation for the next research on Uyghur named entity recognition, and it can also play a positive role in the further research of Machine Translation, information extraction, syntactic analysis, semantic analysis and so on.

## 6. Acknowledgments

This work has been supported as part of the NSFC (61462083, 61762084, 61463048, 61262060 ), 973 Program (2014cb340506), and 2017YFB1002103, ZDI135-54.

## 7. References

Coates-Stephens S. 1992. The analysis and acquisition of proper names for the understanding of free text. Computers & the Humanities, 26(5-6), 441-456.

Thielen C. 1999. An approach to proper name tagging for german.

Sundheim B M. 1996. Overview of results of the MUC-6 evaluation. A Workshop on Held at Vienna, Virginia: May (pp.423-442). Association for Computational Linguistics.

Sang E F T K, Meulder F D. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Conference on Natural Language Learning at Hlt-Naacl(Vol.21, pp.142-147). Association for Computational Linguistics.

Walker C, Strassel S, Medero J, Maeda K. 2006. Ace 2005 multilingual training corpus. Progress of Theoretical Physics Supplement, 110(110), 261-276.

Polifroni J, Kiss I, Adler M. 2010. Bootstrapping Named Entity Extraction for the Creation of Mobile Services. International Conference on Language Resources and Evaluation, Lrec 2010, 17-23 May 2010, Valletta, Malta. DBLP.

Savary A, Waszczuk J, Przepiórkowski A. 2010. Towards the Annotation of Named Entities in the National Corpus of Polish. International Conference on Language Resources and Evaluation, Lrec 2010, 17-23 May 2010, Valletta, Malta. DBLP.

Desmet B, Hoste V. 2013. Towards a Balanced Named Entity Corpus for Dutch. International Conference on Language Resources and Evaluation, Lrec 2010, 17-23 May 2010, Valletta, Malta (pp.535-541). DBLP.

Rozi A, Zong C, Mamateli G, Mahmut R, Hamdulla A. 2013. Approach to recognizing uyghur names based on conditional random fields. Journal of Tsinghua University, 53(6), 873-877.

Tashpolat N, Wang K, Askar H, Palidan T. 2017. Combination of statistical and rule-based approaches for uyghur person name recognition. Zidonghua Xuebao/acta Automatica Sinica, 43(4), 653-664.

Adila Ahmat, Feng Xiangping. 2017. Cognition of Uyghur musical named entity based on condition random Field. [j]. Intelligent Computer and Applications.

Arkin M, Hamdulla A, Tursun D. 2013. Recognition of uyghur place names based on rules. Communications Technology.

Maihefureti, MiriguRouzi, MaierhabaAili, TuergenYibulayin, Department T A. 2014. Uyghur organization name recognition based on syntactic and semantic knowledge. Computer Engineering & Design.

Mahmoud A, Yusuf H, Zhang J, Zong C, Hamdulla A. 2017. Name recognition in the uyghur language based on fuzzy matching and syllable-character conversion. Qinghua Daxue Xuebao/journal of Tsinghua University, 57(2), 188-196.

Yang Y, Xu C, Gulimire A. 2011.uyghur named entity identification basing on maximum entropy model[C]// 3rd International Conference on Information Technology and Computer Science.

Jiazheng L I, Liu K, MairehabaAili, Yajuan L V, Liu Q, TuergenYibulayin. 2011. Recognition and translation for chinese names in uighur language. Journal of Chinese Information Processing, 25(4), 82-87.

Abiderexiti K, Maimaiti M, Yibulayin T, Wumaier A. 2017. Annotation schemes for constructing Uyghur named entity relation corpus. International Conference on Asian Language Processing. IEEE.

Ramshaw L A, Marcus M P. 1995. Text chunking using transformation-based learning. Text Speech & Language Technology, 11, 82--94.