Semi-Automatic Corpus Expansion for Uyghur Named Entity Relation based on a Hybrid Method

Kahaerjiang Abiderexiti[1,2] , Ayiguli Halike[1, 2], Maihemuti Maimaiti[1,2], Aishan Wumaier[1,2] , Wanglulu[1,2], Tuergen Yibulayin[1,2]

[1] School of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang 830046, China
[2] Xinjiang Laboratory of Multi-Language Information Technology, Urumqi, Xinjiang 830046, China
kaharjan@xju.edu.cn, 1506867752@qq.com, mahmutjan@xju.edu.cn,
hasan1479@xju.edu.cn, turgun@xju.edu.cn

Abstract

In order to address the issues that in Uyghur lack of relation extraction method and small size of relation annotated data, we present a semi-automatic way to expand existing Uyghur named entity relation corpus. Our method is based on Conditional Random Fields followed by rules. We integrate our relation extraction method into our annotation tool, with the help of human correction, we expanded existing corpus size by three times.

Keywords: Uyghur, named entity relation, conditional random field

1. Extended Abstract

Relation extraction is the prerequisite task of recognizing and characterizing a particular relationship between two or more entities in text. Depending on the languages in which annotated named entity relation corpora are available, relation extraction has been studied using different machine learning methods, including supervised, semi-supervised and even unsupervised method. Those studies focus on English language and other resources rich languages. However, relation extraction in Uyghur language, which is ethnic minority language that wildly used in Xinjiang Uyghur Autonomous Region of China, there are two problems: 1) there are no studies have reported regarding with relation extraction method. 2) the existing annotated Uyghur named entity relation corpus size is relatively small. To address these issues, we utilized the existing Uyghur named entity relation annotated corpus which only contains a small amount of annotated news articles to propose a hybrid semi-automatic method of expanding existing annotated corpus. Our method is based on Conditional Random Fields (CRFs) followed by some rule based post processing and manual correction. In this way, we expanded the corpus size by three times than the existing one.

2. Introduction

Relation extraction is the most important task in natural language processing, especially in the field of information retrieval, knowledge graph. The aim of relation extraction tasks is recognizing and characterizing a particular relationship between two or more entities in text automatically. There are several methods in relation extraction including supervised methods, semi-supervised methods, distant supervised methods and unsupervised methods. Supervise and semi-supervised methods require human annotated data depending on which methods are applied. Usually, supervised methods require more annotated data than semi-supervised methods. Unsupervised may not require annotated data but still need large amount of unannotated data. For Uyghur language, one of the official languages in Xinjiang Uyghur Autonomous Region in China, both annotated and unannotated data are scarce. However, relation extraction requires a certain amount of annotated corpus especially in supervised learning. The size of Uyghur named entity relation annotated data which is available on the Internet is relatively small. It is difficult to train Uyghur relation extractor using this small data. The data size has become one of the major limitations in studying relation extraction in Uyghur. And also there is not any report about relation extraction in Uyghur language. To solve these problems, we proposed the hybrid semi-automatic method that expanding existing annotated corpus based on conditional random fields (CRFs).

In this paper, first we describe related work in Uyghur corpus construction and relation extraction in other languages. Then describe our method which aims to expending existing corpus size by relation extraction. Finally, we show our results and discuss further improvements.

3. Related Work

Entity relationships are the key to building a knowledge graph, there are many methods for relation extraction. In the traditional approaches, entity recognition is considered as a predecessor step in a pipeline for relation extraction (Zelenko,etal.2003). However, it is ignored the dependence between the two tasks. and (Li Y, et al.2011) research entity relation descriptor based on linear-chain CRFs, and that reduce the space of possible label sequences and introduce long-range features. Recently, neural network based methods have become popular in natural language processing. In the relation extraction, there are also several methods which are based on neural network and also methods with using joint extraction with named entity. R kai, W Shi-Wen(2016) proposes an abbreviation disambiguation method based on the convolutional neural network (CNN) to solve the abbreviation disambiguation problem in the biomedical field when no labelled corpus exists and obtained an average of 90.1% accuracy. However, in this way the named entity recognition accuracy affects the relation extraction. Join extraction of entities and relations is the new way to address this

problem. (Zheng, et al.2017) use the novel tagging scheme whose annotations can be converted joint extraction task and study different end-to-end models to extract entities and relations . (Miwa M and Bansal M, 2016) propose end-to-end relation extraction method used BI_LSTM, and this model can extract jointly both named entities and relations with shared parameters.(Zhang M, Zhang Y,2017) also proposed global optimized neural model ,and achieving the best performances on two standard benchmarks.In Uyghur language processing, there are some works about constructing corpus, particularly, to solve shortness of Uyghur named entity and named entity relation corpus, (Kahaerjiang Abiderixiti, et.al, 2017) proposed the method for construction Uyghur named entity and relation corpus and release small size of annotated corpus. Wushouer J, et al. constructed "contemporary Uyghur grammatical Information Dictionary", which is provided a large amount of grammatical information and collocation features, and it is the basic resources of NLP. There is also a research about building Uyghur Dependency Treebank which is built from a public reading corpora (Aili M),and it is also important for linguistic researches.

4. CRFs Model

The task of relation extraction can be seen as a sequence labeling problem. For the small amount of data, conditional random fields (CRFs) model would be more effective. So we choose CRFs model as the main model expending our corpus by the help of relation extraction model. CRFs is one of the commonly used algorithms in natural language processing in recent years, which combine the maximum entropy and hidden Markov model, which is a typical non-directional pattern model of discriminant probability. The CRFs attempt to model the conditional probability of multiple variables after a given observed value. $X = \{x_1, x_2, x_3 \dots \dots x_n\}$ are the observed sequence, $Y = \{y_1, y_2 \dots \dots y_n\}$ are the corresponding tag sequences. Construction of conditional probability model is the goal of CRFs (Maimaiti, M.). The definition of CRF model (Lafferty,et al. 2001) is as follows:

Definition: set $G(V, E)$ as a node of V, and E as an undirected graph of a set with no edges.

$Y = \{Y_v | v \in V\}$, Each node in V corresponds to a random variableY_v, whose value range is a possible set of tags. If an observed sequence X for conditions, each random variable satisfy Markov characteristics as follows:

$$P(Y_V \backslash X, w \neq v) = P(Y_V \backslash X, Y_w, w \sim v) \quad (1)$$

Where W: V, denotes that two nodes are adjacent nodes in graph G. Then, (X, Y) as a conditional random field. The CRFs model calculates the conditional probability model of the output node as the condition of given input nodes. For a chain with parameter $\theta = \theta_1 \theta_2 \dots \theta_k$,the conditional probability of a state sequence obtained for a given sequence x is defined as:

$$P_\theta(Y|X) = \frac{1}{Z_x}\{\sum_{n=1}^{N} \sum_k f(y_{n-1}, y_n, X, n)\} \quad (2)$$

Among them, the denominator Z_x normalized factor:

It enables that the sum of the probabilities of all possible state sequence of the given input to be 1. $f_k(y_{n-1}, y_n, X, n)$for the whole observing sequence X, the feature functions, which are located at n and n-1, may be 0, 1, or any real number. In most cases, the characteristic function is a binary representation function; When the characteristic condition is satisfied, the value is 1, otherwise it is 0. The definition of characteristic function as shown below:

$$f(y_i, x, i) = \begin{cases} 1 & if\ y_{i-1} = y` and\ y_i = y \\ 0 & other \end{cases} \quad (3)$$

$$g(y_i, x, i) = \begin{cases} 1 & if\ x_i = x\ and\ y_i = y \\ 0 & other \end{cases} \quad (4)$$

$\theta = \theta_1 \theta_2 \dots \theta_k$ is the corresponding weight for the feature function. For X, in the next step is to search the $Y* = \text{argmax}\ P(X\backslash Y)$ with the highest probability.

5. Uyghur Relation Extraction Model

On the basis of analysis in the grammatical and semantic features of Uyghur named entity relation, we firstly use the CRfs model to study Uyghur relation extraction method. The reason of using CRFs is that it is better neural network models when the data size is relatively small. In the design of features, we use different Uyghur language features such as words, syllables, part-of-speech tagging and distributed word vector representations.

5.1 Task Definition

As we said above , relation extraction task can be seen as sequences labeling problem. As shown on figure 1, annotation for relation extraction from raw texts. it only consists of relation extraction tags, which recognizes valid relations over entity pairs. According to the characteristics and difficulties of Uyghur Named Entity and relations, we construct feature sets between the different entity categories. The features of our method are divided into word related features and the dictionary features. Firstly, word related features include Uyghur words itself, part-of-speech tagging, syllable, word length and syllable length, etc. Because Uyghur words stemming is complicated and there are no good stemming tools publicly available., so we didn't use stem characteristics. Secondly, the dictionary features, It`s feature include common dictionaries, person name dictionaries, place name dictionaries, organization name dictionaries, and the similarity dictionaries based on word vectors, etc. The following Table 1 shown as, the feature information of a every word in the Uyghur sentence.

5.2 Feature Template

The influence of combining different features on the named entity relations extraction can't be ignored. Therefore, the selection of feature templates plays an extremely important role in relation extraction. Named entity relations extraction need to consider the context, whereas the CRFs model synthesizes contextual information as well as external features. Named entity relation recognition needs to consider the context, whereas the CRFs model synthesizes contextual information as well as external features.

In this paper, we use CRFSharp open source tools to build Uyghur named entity relation recognition model, the use of defined characteristics acquired template feature to learn. In the model, not only the atomic feature (unary feature) template, but also the composite feature template has to be defined. The definition of feature templates common to throughout in this paper is given in Table 2.

words	$feature_1$	$feature_2$	$feature_{...}$	$feature_n$	Final Tag
شىنجاڭ	Arg1	Arg2	…….	Arg_n	B_Org-Aff.Employment_1
ئۇنۇۋېرسىتتى	Arg1	Arg2	…….	Arg_n	E_Org-Aff.Employment_1
مەكتەپ	Arg1	Arg2	…….	Arg_n	O
مۇدىرى	Arg1	Arg2	…….	Arg_n	O
ۋەلى	Arg1	Arg2	…….	Arg_n	B_Org-Aff.Employment_2
ياقۇپ	Arg1	Arg2	…….	Arg_n	E_Org-Aff.Employment_2
.	Arg1	Arg2	…….	Arg_n	O

Table 1: Sequence labeling tags for relation extraction task

Feature type	Template	Meaning
Atomic feature	$w_i(-2 \le i \le 2, i \in Z)$	The current word w_i and its upper and lower two window words, the word's window size is 5
	$F_i(-1 \le i \le 1, i \in Z)$	The characteristics of the current word F_i and the words of its upper and lower windows, ie the window size is 3
Composite feature	$w_{i-1} \mid w_i(0 \le i \le 1, i \in Z)$	The combination of the current word and its upper words feature
	$F_{i-1} \mid F_i((0 \le i \le 1, i \in Z))$	The combination feature of the current and upper words
	$w_i \mid F_i(i = 0)$	The current word and its combination features
	$F_{i-1} \mid F_i \mid F_{i+1}(i = 0)$	The characteristics of the current word and compound features of one window`s upper and below

Table 2: feature template

In Table 2, w represents the first column of the corpus, that is, a column of words, and F denotes other characteristic columns except words; in which the $F_{i-1}|F_i|F_{i+1}(i=0)$ in the composite feature represents a combination of three features, and the other three features represent the binary feature combinations.

5.3 Rules

The relation is different from the Named Entity. And Uyghur relation extraction is more difficult and more challenging. Because the annotated data size is small and Uyghur language is morphologically complex. We simplified the task and assume that the named entity is given. So we just tagged the relationship between the named entities represented in a sentence. However, the result still did not help much annotator. So after relation extraction based on CRFs. we use some rules to posted it the extraction result. Relation annotation has some rules. What we are considering here is the rules for the annotated relations.

5.3.1 Physical.Located

Physical location relations: The first argument of this relationship must be a person. The second argument can be for facilities (FAC), location (LOC) and geographical social and political entities (GPE) shown by Table 3.

Relation type	Arg1	Arg2
Physical.located	PER	FAC,LOC,GPE

Table 3: Candidate Arguments for Physical.located

In the example below, گۈزەلنۇر (Guzalnur) is the Arg1 (PER) and it is must be a person and شاڭخەيدە (in Shanghai) is Arg2 and it is belonging to LOC.

گۈزەلنۇر شاڭخەيدە ئوقۇۋاتىدۇ .

Guzalnur is studding in Shanghai

5.3.2 Physical.Near

The rules for Physical.near relation is also shown in Table 4.

Relation type	Arg1	Arg2
Physical.Near	PER,FAC, GPE,LOC	FAC,LOC,GPE

Table 4: Candidate Arguments for Physical.Near

In the example below, قەشقەر ۋىلايىتى (Kashagar prefecture) is the Arg1 (GPE), ئاقسۇ ۋىلايىتىنىڭ (Aksu prefecture)is the Arg2 (GPE).

قەشقەر ۋىلايىتى ئاقسۇ ۋىلايىتىنىڭ جەنۇبىغا جايلاشقان .

Kashgar prefecture is located the sought of Aksu prefecture.

As described above, all the subtypes of named entity relations have certain rules, which rules similar to previous example. And all types and subtypes are classified as follows:

I. GenPart-Whole (Geographical, Subsidiary)
II. Personal-Social (Business, Family, Role, other)
III. Physical (Located, Near)
IV.ORG-Affiliation (Employment, investor-Shareholder, Student-Alum, Owner, Founder)
V. General-Affiliation
(PersonAge, Organizationwebsite).

6. Results

After we add rules for post edit which descried section 5, our suggestion for relation annotation is improved. We got positive feedback from our human annotator.

As the result, we have mainly completed the following two tasks. The first one is, we integrate our relation extraction model into our annotation tool.

The second one is, when annotating the named entity relationship corpus, our model provides a suggestion for annotation, thereby alleviate the burden of human annotators. . Thus, the scale of our annotated corpus is increasing rapidly. The annotation result of this tool is shown as below:

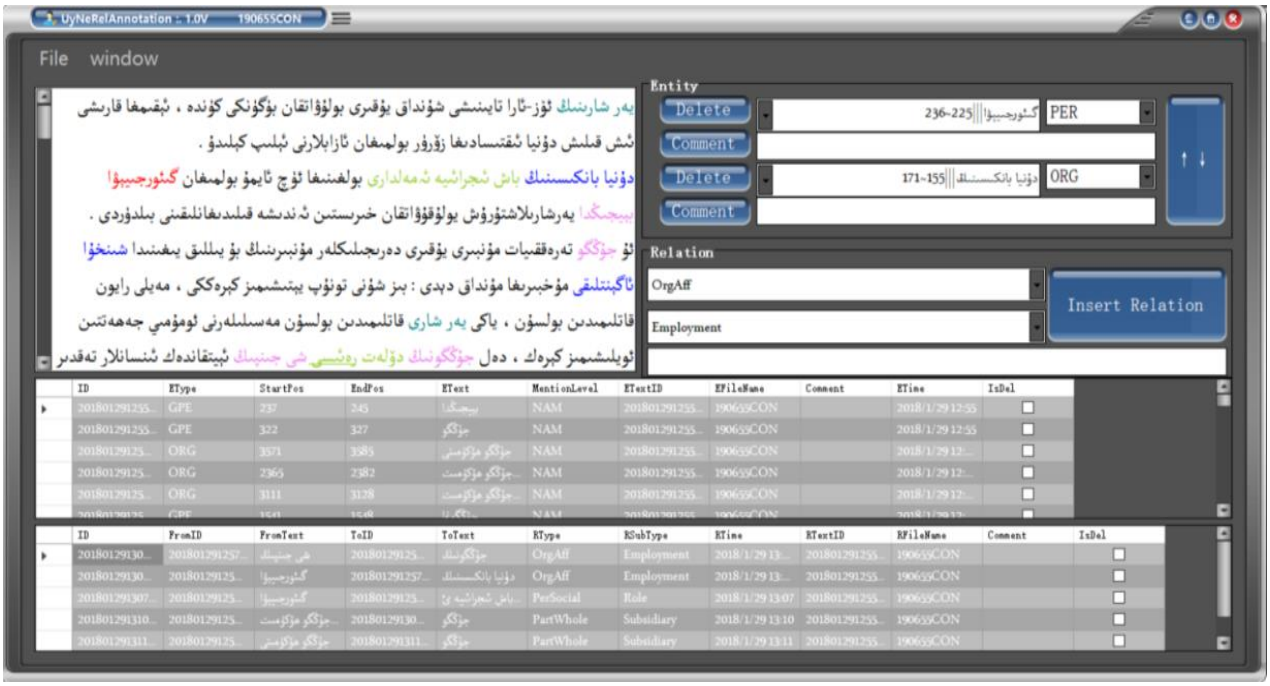

Figure 1: Interface of the UyNeRel Annotation Software

We estimated that our human annotation speed is speedup more than three times. Two annotators who are already familiar with relation annotation annotate 21 documents per day before integrating our relation extraction model into our relation extraction tool. After we add our model to annotation tool their annotation speed is increased and they have annotated 62 documents per day on average. In this semi-automatic way, the existing 500 documents increased to 1500 in 77 days in average. As the result the relation annotation corpus size is expanded by three times in this semi-automatic way.

7. Conclusion

In this study, we described our work on expending Uyghur Named Entity Relation, the main purpose of this article is to provide the extension annotated corpus which is needed in the study of automatic relation extraction. In the immediate future, we plan to focus on the relation extraction task based on neural network in Uyghur.

8. Acknowledgements

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

9. References

Guodong, Z., Jian, S., Jie, Z., & Min, Z. (2002). Exploring Various Knowledge in Relation Extraction. ACL 2005, Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, Usa (419--444). DBLP.

Zheng, S., Wang, F., Bao, H., Hao, Y., Zhou, P., & Xu, B. (2017). Joint extraction of entities and relations based on a novel tagging scheme. 1227-1236.

Wingendcr, M. (2007). Standardisation tendencies in an expanded europe - a corpus-based study of the anglicisms in polish. Welt der Slaven-Halbjahresschrift fur Slavistik, 52(1), 1-20.

Rimkus, C. D. M., Junqueira, T. D. F., Callegaro, D., & Leite, C. D. C. (2013). Segmented corpus callosum diffusivity correlates with the expanded disability status scale score in the early stages of relapsing-remitting multiple sclerosis. Clinics, 68(8), 1115-1120.

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.

Wushouer, J., Abulizi, W., Abiderexiti, K., Yibulayin, T., Aili, M., & Maimaitimin, S. (2015). Building Contemporary Uyghur Grammatical Information Dictionary. Revised Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure (pp.137-144). Springer-Verlag New York, Inc.

Aili, M., Xialifu, A., Maihefureti, & Maimaitimin, S. (2015). Building Uyghur Dependency Treebank: Design Principles, Annotation Schema and Tools. Worldwide

Language Service Infrastructure. Springer International Publishing.

Maimaiti, M., Wumaier, A., Abiderexiti, K., & Yibulayin, T. (2017). Bidirectional long short-term memory network with a conditional random field layer for uyghur part-of-speech tagging. Information, 8(4), 157.

Li, Y., Jiang, J., Hai, L., Ming, K., & Chai, A. (2011). Extracting relation descriptors with conditional random fields. Asian Federation of Natural Language Processing, 392-400.

Zhang, M., Zhang, Y., & Fu, G. (2017). End-to-End Neural Relation Extraction with Global Optimization. Conference on Empirical Methods in Natural Language Processing (pp.1730-1740).

Miwa, M., & Bansal, M. (2016). End-to-end relation extraction using lstms on sequences and tree structures.