International Journal of Database Theory and Application
Vol.9, No.1 (2016), pp.209-216
http://dx.doi.org/10.14257/ijdta.2016.9.1.18

Research on Method for Uyghur Temporal Word Recognition

Azragul[1,2], Alim Murat[1,2] and Yusup Abaydula[2]

[1]*The Xinjiang Technical Institute of Physics & Chemistry CAS, Urumqi, Xinjiang, 830011, China*
[2]*School of Computer Science & Technology Xinjiang Normal University, Urumqi, Xinjiang, 830054, China*
Azragul2010@126.com

Abstract

Researches concerning to the temporal expressions in minority languages, particularly the Uyghur temporal word recognition, has not previously been conducted. this paper stated for the first time the relevant progresses and significances on domestic and overseas researches. We analyzed the formation of simple and compound temporal words in modern Uyghur language. We discussed the constitutive rule for Uyghur temporal expressions and put forward a new rule for combined temporal expression templates on the basis of dictionary and regular expressions. We then designed a suitable recognition algorithm and implemented an extraction system for the Uyghur temporal expressions. In the end, the feasibility and usability of the complete recognition method and results were discussed.

Keywords*: Modern Uyghur Language Temporal Word, Method for Recognition*

1. Introduction

Temporal information is an important message in natural language, and a major part of an event. Study indicates that its proportion in the text information is just behind proper noun. In daily life, people always want both content and temporal information that are in the article to be connected, when they read a piece of news, and can understand the beginning, the process, the end and the frequency of event occurrence, moreover, be able to grasp the whole process of event occurrence and understand the ins and outs of event, as an important basis for further decisions through the time information in text. Therefore, temporal information processing is a very important part that in the process of understanding natural language and it have made a great value to the fields of that information extraction, information retrieval, question answering, text digest and data mining.

According to some domestic and overseas researches status, there are two types of main method for temporal expression recognition. One is a machine learning-based sequence labeling approach; the other one is a rule-based approach. In the first method, machine learning-based sequence labeling method, even it has achieved higher recall rate just depending on annotated context, and that it's over dependency on the quality of annotated corpus and making the less use of features relative stability of temporal expression recognition, for which ,it can be less used in temporal expression recognition; in the rule-based approach, despite the fact that it has made a good use of features relative stability of temporal expression recognition, and built a complete rule of temporal expression recognition , however, this kind of rule often has an over-coarse grained disadvantage and does not recognize the very fine temporal expression, meanwhile, there are plenty of handwritten rules causing a lot workload.

At present, research on English and Chinese temporal expression recognition has been reached an in-depth stage, and many rules which are largely different from Uyghur

ISSN: 2005-4270 IJDTA

culturally and linguistically has arisen, besides, these rules can not directly be used for Uyghur temporal expression recognition, so that it has become a hard task for Uyghur language. Hence, that build a rule by observing and extracting all features of temporal expression in Uyghur language is currently the best solution which has got a quite reliability, efficiency and scientificity.

2. Analysis on Temporal Words in Modern Uyghur Language

2.1 Morphology of Temporal Word

Uyghur language, a very developed language which has a various morphological system and it always adopts various grammatical forms to express the relevance between actions or states and speech time that verbs refer to as a tense.

From the point of relation between verbs and speech time, Uyghur language can be divided into three tense as following: present, past and future. For example:

مەن بازارغا بارىمەن I am going to the market.

In this sentence, as بارىمەن "going to go" only expressed a future tense, and it can't be able to determine an exact time. Moreover, when it is said that

مەن بازارغا ئەتە بارىمەن I am going to the market tomorrow.

In this sentence, ئەتە بارىمەن "going to the market tomorrow" not only expressed a future tense, but including a word which can be formalized and determining an exact time. In its morphology, verbs roughly show that action or state may possibly happen in the past, now or in the future, and could not have said any concept of time, at the same time, can't show the agglutinating characteristics of Uyghur language.

2.2 Form of Temporal Word

Temporal word refers to as a concept of time which is demonstrated by the meaning of word itself and that word represents a time period is called temporal word.

1) Temporal Noun: (Day) كۈن, (Month) ئاي, (Year) يىل, (Hour) سائەت, (Minute) مىنۇت, (Second) سېكونت, (century) ئەسىر, (quarter) پەسىل, (week) ھەپتە etc. Uyghur time nouns have morphological changes in person, so that these time nouns appear different forms in the sentences.
2) Time Adverb: (sometime) گاھ, (always) ھەمىشە, (from now on) ئەمدى, (awhile) بىردەم, (often) ھامان, (permanent) مەڭگۈ, (usually) دائىم etc. Uyghur time adverbs generally do not have morphological changes, but there are very few adverbs have a less meaning of time, when they connected with affix.
3) Compound Temporal Word: (today) بۈگۈن, (this year) بۇيىل, (from tomorrow) ئەتىدىن باشلاپ, (year from 2012 to 2014) 2012 - يىلغىچە - 2014 يىلدىن, (tomorrow at noon) ئەتە چۈش, (till tomorrow) ئەتىگە قەدەر etc.

Above three notation of time in Uyghur respectively, it all can be able to express accurate temporal information. So that is to say, when using morphological form can only shows that action or states occurs in the past, now or in the future. In order to indicate the accurate time, we need to add corresponding temporal words in a sentence.

About research on temporal words in English and Chinese, having a very long systematic theory, in this way, their work on temporal information extraction mostly focused on temporal word recognition. On the contrary, Uyghur language needs to be studied further in this field. Hence, this paper aims at recognizing the obvious temporal words in Uyghur text.

3. Rule-based Temporal Word Recognition

Rule-based approach is one of the important methods in Natural Language Processing. Reason that comparatively formalized features of temporal word in Uyghur, for which rule-based approach is more suitable for this task. In Uyghur language, research on entity recognition, even the studies on temporal word recognition is at the first step now. So this article mainly focused on a rule-based approach, and designed Uyghur temporal word recognition system by using a regular expression and dictionary combined method.

3.1 Regular Expression-Based Rule

Regular expression is a rule, using single or multiple strings to describe, match a series string in line with a syntactic rules. There are very simple and flexible time information in Uyghur language, which is Time, Date, Duration and TN (Temporal Noun) etc. in some cases, besides these simple time words, Time Adverbs, TLN (Time Locality Noun) and P (Prepositions), can be combined into complex temporal word.

In the form of Uyghur temporal information, the time consists of (Hour) سىكونت, (Minute) مىنۇت and (Second) سائەت. The date consists of (Year) يىل, (Month) ئاي, (Day) كۈن. Apparently, both time and date arrange in left to right order. About duration, it basically has a rule to follow and its general format is: [Number + Quantifier + TN]. For time word, find a perfect match with the text using time word dictionary. So as easily to express this temporal information and make it easier when it is being converted to regular expression, we divided temporal information into two templates under the rules and characteristics of Uyghur language in this paper. One is a simple temporal word; another is a compound temporal word. Templates are as follows:

Table 1. Template for Simple Temporal Word

Temporal Information Types	Examples(En)	Examples(Uy)
Time	Seven past thirty; 7:30	سائەت 7دىن 30 مىنۇت؛ سائەت يەتتىدىن ئوتتۇز مىنۇت
Date	June 1st, 2013	2013-يىل 6-ئاينىڭ 1-كۈنى؛ 2013-يىل 1-ئىيۇل
Temporal Noun	Today; Winter; Morning	بۈگۈن؛ قىش؛ ئەتىگەن
Duration	[Three|3] minutes; [Two|2] months; [Five|5] years	[ئۈچ|3] مىنۇت؛[ئىككى|2] ئاي [بەش|5] يىل

Table 2. Templates for Complex Temporal Word

Temporal Information Types	Examples (En)	Examples(Uy)
TN + Time	This morning [ten|10] o'clock	ئەتىگەن\TN سائەت [ئون|10]\Time
Date + TN	The summer of 2013	2013 ـ يىل\Date ئەتىياز\TN
TN + Date	The last June 11th	ئۆتكەن يىلى\TN 11 ـ ئىيۇن\Date
Date + Time	Feb. 14th 8 past 45	2 ـ ئاينىڭ 14 ـ كۈنى\ Date سائەت8دىن 45\Time
Date + TN + Time	Sept. 1st, 2013 AM 10:30	2013 ـ يىل9 ـ ئاينىڭ1 ـ كۈنى\Date چۈشتىن بۇرۇن\TN سائەت 10دىن 30مىنۇت\Time
TN + TN	Last night; Monday morning	تۈنۈگۈن\TN كەچ\TN؛ دۈشەنبە\TN ئەتىگەن\TN
TN + Duration	The past year; The next [12|twelve] days	ئالدىنقى\TN [1|بىر] يىل\Duration؛ كەلگۈسى\TN[12|ئون ئىككى]كۈن\Duration
[Time|Date|TN|Duration]+TLN	Before [eight|8] o'clock; After Sept. 1st, 2013; [five|5] days ago	سائەت[8|سەككىز]دىن\Time بۇرۇن\TLN؛2013

		ـ يىل9 ـ ئاينىڭ 1 ـ كۈنىدىن\Date كېيىن\TLN؛ [بەش|5] كۈن\Duration ئىلگىرى\TLN
[Date|Time|TN]+P	By the Sept. 1st, 2013; By the [ten past thirty|10:30]; It will be today	2013 ـ يىل9 ـ ئاينىڭ1 ـ كۈنى\Dateگىچە\P؛ سائەت10يېرىم\Timeغىچە\P؛بۈگۈنگە\TN قەدەر\P
[Date|Time|TN]+P+TLN	Since Aug.1st; By 10:30AM;Since the Spring Festival	8 ـ ئاينىڭ 1 ـ كۈنى\Dateدىن\P بۇيان\TLN؛سائەت10يېرىم\Time دىن\P ئىلگىرى\TLN؛باھار بايرىمى\TN دىن\P بۇيان\TLN
[Time|Date|TN|Duration]+P+ [Time|Date|TN|Duration]	From Oct.15th,2012 to the early July of 2013; From last summer to the June of this year	2012 ـ يىل10 ـ ئاينىڭ15 ـ كۈن\Date بىدىن\P 2013 ـ يىل7 ـ ئاينىڭ\Date بېشىغىچە؛ئۆتكەن يىل ياز\TNدىن تارتىپ\P بۇيىل6 ـ ئايغىچە\Date
.....	.....	.....

Above-mentioned various templates are that basic form of Uyghur temporal information, have a certain fixed format and rules to follow. Eventually, these templates can be transformed into regular expression which could have searched a temporal word in order to match in each sentence.

3.2 Common Temporal Noun dictionary

Temporal noun as an individual word generally refers to as a word expressing temporal information. For example:ئەتىگەن (Morning), ھارپا (Eve), تاڭ سەھەر (Dawn), گۈگۈم (Evening), دۈشەنبە (Monday) *etc*.

In regular expression-based rules, though different templates for temporal word have given, but the fact that there are small numbers of temporal noun, word that is a formalized and most frequently appeared in certain point, and it can also compose a compound temporal word by combining with simple temporal word in which of Time, Date and Duration. Therefore, collect some common temporal noun and built a dictionary is another major part of this work.

Table 3. Most Frequent Uyghur Temporal Noun

Uyghur Temporal Noun	Example(En)
بۈگۈن	Today
ئۆلۈشكۈن	The day before
بۈرناكۈن	The day before
ئەتە	Tomorrow
ئۆگۈن	The day after
تۈنۈگۈن	Yesterday
دۈشەنبە	Monday
سەيشەنبە	Tuesday
قىش	Winter
ياز	Summer
.....	.....

Both regular expression and dictionary combined rule we mentioned above, not only can recognize a temporal word without any affixes connection and compound temporal word, but a temporal word can be affixed. As for temporal noun dictionary, it is mainly about temporal noun stem, in order to save memory space. In terms of affixed temporal word, using positive maximum match method to determine boundary between temporal word and affix, that is to say, matching the longest string as a temporal word.

4. Rule-based Temporal Word Recognition Algorithm

In this paper, method for recognizing Uyghur temporal word mainly adopts two-step strategy. First, recognize simple temporal words based upon templates. Such as: Time, Date and Duration. Second, it will begin matching each simple temporal word by using templates for complex temporal word, just after it had finished recognizing simple temporal word. If a two adjacent simple temporal word and one of templates for complex temporal word have a match, then match is successful, thus, the two adjacent simple temporal words can be a new compound temporal word. Otherwise, there is no any match so that match is fail and it will skip to next sentence. Flow process is as follows:

1) Input the text;
2) Scanning the whole sentences, and recognizing basic simple temporal words such as Date, Time and Duration;
3) Scanning simple temporal words in sentences, whether it can match with template for complex temporal word?
   Y: Merge simple temporal words into a complex temporal word, and turn to step (3);
   N: turn to step (4);
4) Output the final Uyghur temporal word;
5) Whether is the end of text?
   Y: Turn to step (6);
   N: Scan next sentence, and turn to step (2);
6) End

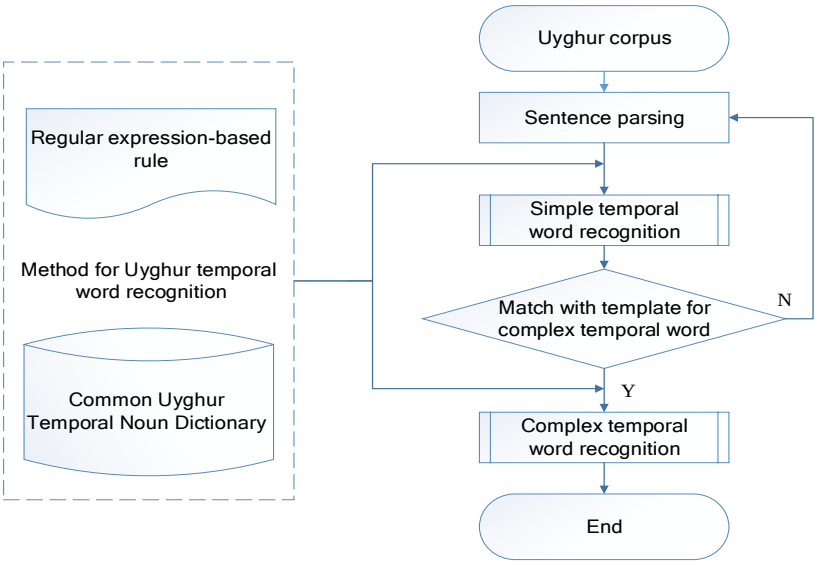

Figure 1. Flow Process of Uyghur Temporal Word Recognition

5. Experiment and Analysis

5.1 Corpus and Evaluation

In this article, Experimental corpus that is provided by Key Laboratory for network security and public analysis in Xinjiang Normal University, which contains the whole text of semi-annual daily half-hour broadcast of “News Network” and “Xinjiang News”, and size of 6.74MB.

As Uyghur temporal word recognition is a subtask of named entity recognition, in order to evaluate the performance of recognition using the most classic “PRF evaluation method” in the named entity recognition field. Calculation formula is as follows:

$$Precision\,(P) = \frac{\text{Total temporal words that are only correctly recognized}}{\text{Total temporal words that are recognized}} \times 100\%$$

$$Recall\,(R) = \frac{\text{Total temporal words that are only correctly recognized}}{\text{Total standard temporal words in corpus}} \times 100\%$$

$$F\text{-}Score = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times (P + R)} \times 100\%$$

5.2 Results

Table 4. Corpus statistics

Files	The whole temporal words	The recognized temporal words (Correct)	The recognized temporal words (Incorrect)	The unrecognized temporal words
2243	3407	3042	75	365

Table 5. Evaluation results

Precision	Recall	F-Score
97.5%	87%	91.95%

The results in Table 6 showed that both regular expression and dictionary combined approach in Uyghur temporal word recognition in which precision is about 97.5% and recall is 87%, and the rules and templates fully coincide with the grammatical form of Uyghur temporal word.

However, as the system returns a higher precision, but it couldn't have a high coverage rate of each temporal word in the whole corpus, so that it is appeared that recall rate slightly lower than precision. There are following aspect cause not high recall rates:

1) Because of incomplete Uyghur temporal information types in two different templates, therefore it cannot able to cover all regular expression rules in Uyghur temporal word.
2) There are some incorrect Uyghur characters which do not follow orthographic rule of Uyghur language, and unable to match templates for Uyghur temporal word in corpus, so that it cause a bit increases of the unrecognized temporal words.
3) There are not enough Uyghur temporal nouns in dictionary.

Overall, the rule-based Uyghur temporal word recognition, its evaluating result is over 85% whether in precision or in recall, so which reflects rules and algorithms that in this article proved its feasibility and essential applied value in minority language processing. Mean while, the comprehensive performance index that F-score reaches 92%, this also shows that our method has a rather high efficiency in overall test.

6. Conclusion

This article carried out certain research and work on Uyghur temporal word recognition through analyzing the morphology of temporal word and its formation, then presented an approach that contains regular expression rule and temporal noun dictionary, finally implemented Uyghur temporal word recognition system. Test result showed that Uyghur temporal word recognition system based on both regular expression and temporal noun dictionary achieved higher rate in precision and recall respectively 97.5% and 87%.

Acknowledgements

Funding for this research was provided by from Natural science foundation of the Xinjiang Uyghur autonomous region (project number: 2014211A045); Philosophy and social science research of the Xinjiang Uyghur autonomous region planning fund project (project number: 14CYY093); Ministry of education of humanities and social science in general project (project number: 14YJC740001); The Xinjiang Uyghur autonomous region research start fund for young teachers in university's scientific research plan (project number: 20140706213103147); Key project of national natural science foundation of China (project number: 61132009); The national natural science fund project (project number: 61262066); Key national social science fund (project number: 14AZD11).

References

[1] P. Yuequn, "Research on temporal Information Recognition and Normalization", Harbin Institute of Technology, **(2006)**.
[2] W. Yun, "Chinese Temporal Information Extraction in Financial Field", Tsinghua University, **(2004)**.
[3] W. Tong, "Research on Chinese Time Expression Recognition", Fudan University, **(2010)**.
[4] L. Junchan, T. Hongye and W. Fenge, "Recognition of Temporal Expression and their Types in Chinese", Computer Scince, vol. 39, no. 11A, **(2012)**, pp. 191-194.
[5] L. Jie, "The express of Past Tense Contrast between Uyghur and Chinese", Xinjiang University, **(2007)**.
[6] H. Tomur, "Modern Uyghur Grammar", Ethnic Publishing House, **(1987)**, pp. 325-370.
[7] Y. Lei and C. Junyi, "Study on Uyghur Time Adverb", Language Research, no. 10, **(2006)**, pp. 155-156.
[8] W. Fuling, "Comparative Study on Temporal Expression Rule between Uyghur and Chinese", Language and Translation, no. 4, **(1994)**, pp. 27-31.
[9] A. Kadir, "Comparative analysis on Temporal Noun between Uyghur and Chinese", Journal of Changchun University, vol. 23, no. 7, **(2013)**, pp. 836-838.
[10] H. Ruifang, Q. Bing, L. Ting, P. Yueqin and L. Sheng, "Recognizing the Extent of Chinese Time Expression Based on the Dependency Parsing and Error-Driven Learning", Journal of Chinese Information Processing, vol. 21, **(2007)**, pp. 36-40.
[11] Z. Guorong, "Research on Temporal Expression of Chinese News", Shanxi University, **(2006)**.
[12] W. Tong, Z. Yaqian, H. Xuanjing and W. Lide, "Chinese Time Expression Recognition Based on Automatically Generated Basic Time Unit Rules", Journal of Chinese Information Processing, vol. 24, no. 4, **(2010)**, pp. 4-10.

Authors

Azragul, She was born on October 12, 1987, in Xinjiang, China. She is a lecturer in School of Computer Science & Technology Xinjiang Normal University, Currently reading a doctorate in Computer Applications Technology at China's Academy of Sciences. Her main research fields are computational linguistics and natural language processing. Email:Azragul2010@126.com

Alim Murat, He was born on October 15, 1988, in Xinjiang, China. He is currently reading a doctorate in Computer Applications Technology at China's Academy of Sciences. Her main research fields are computational linguistics and natural language processing.

Yusup Abaydula, He was born on October 31, 1958, in Xinjiang, China. He is a professor in School of Computer Science & Technology Xinjiang Normal University. His main research fields are computational linguistics and natural language processing. Email:ysp2002@126.com

