J Inf Process Syst, Vol.14, No.4, pp.824~836, August 2018
https://doi.org/10.3745/JIPS.02.0089

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Applying Lexical Semantics to Automatic Extraction of Temporal Expressions in Uyghur

Alim Murat*, Azharjan Yusup*, Zulkar Iskandar*, Azragul Yusup*, and Yusup Abaydulla*

Abstract

The automatic extraction of temporal information from written texts is a key component of question answering and summarization systems and its efficacy in those systems is very decisive if a temporal expression (TE) is successfully extracted. In this paper, three different approaches for TE extraction in Uyghur are developed and analyzed. A novel approach which uses lexical semantics as an additional information is also presented to extend classical approaches which are mainly based on morphology and syntax. We used a manually annotated news dataset labeled with TIMEX3 tags and generated three models with different feature combinations. The experimental results show that the best run achieved 0.87 for Precision, 0.89 for Recall, and 0.88 for F1-Measure in Uyghur TE extraction. From the analysis of the results, we concluded that the application of semantic knowledge resolves ambiguity problem at shallower language analysis and significantly aids the development of more efficient Uyghur TE extraction system.

Keywords

Feature Combination, Lexical Semantics, Morphosyntax, Temporal Expression Extraction, Uyghur

1. Introduction

A temporal expression (TE), also named TIMEX, refers to any natural language phrases that denote temporal information or a temporal unit, such as an interval or a time point. The extracted TE in the text is so beneficial that time related information is considered as a second informative part in the natural text just behind the proper noun and those TEs are always linked together with content of the article for readers to better understand the entire process of the event.

TE extraction can also be adopted to other natural language processing (NLP) areas. These include, but are not limited to, the following. In question answering system, it is very necessary to answer the "when", "who", "what" and "where" kind of questions and is often seen as a basic element to related task [1]. In summarization system, the ability to allocate events in time aids in acquiring better summaries when it focuses on a particular time period [2]. In recent times, TE extraction has also been applied to other domains like medical information processing [3].

Many works have been accomplished and achieved superb results on temporal annotation in English, Spanish, German and Chinese (see Section 2). But there is still a lack of such resources and systems for

Manuscript received February 22, 2017; first revision July 3, 2017; accepted July 4, 2017.
Corresponding Author: Azragul Yusup (azragul2010@126.com)
* School of Computer Science and Technology, Xinjiang Normal University, Urumqi, China (alim.murat@ms.xjb.ac.cn, {azharjan, zulkarjan, azragul2010, ysp2002}@126.com)

Uyghur language, which annotate documents according to the TIMEX3 standard. In addition, most of the generic approaches to TE extraction are based on explicit rule base encoded in the form of patterns and morphosyntactic feature used for statistical model construction. Nonetheless, these approaches often have difficulty in dealing with semantic ambiguity and generalization at language analysis level. Example (1) illustrates the problem of ambiguity by showing an Uyghur word باهار (underlined in sentence) which has two different senses in sentence. In this case, the difficulty arises on how to differentiate semantically ambiguous words and extract the actual TEs from the text.

Example (1)	**باهار** كىتاپ ئوقۇشقا ئامراق.	(Female proper name)
	نۇرغۇنلىغان ھايۋانلار **باھاردا** كۆپىيىدۇ.	(Season, TIMEX31)

In order to accurately extract Uyghur TEs, in this paper, we make a hypothesis that the linguistic expression of time is a semantic phenomenon and hence, TE extraction must be tackled with semantics. Also, Filannino and Nenadic [4] has indicated that WordNet is compatible to a multilingual extension. At this point, lexical semantics for Uyghur TE is ideal to test its viability and practicability in various minority language processing issues. We, therefore, develop a conditional random field (CRF) based statistical model using semantics. This is based on semantic knowledge (lexical semantic network for Uyghur) plus morphosyntactic knowledge. In so doing, we extract TEs in a precise manner and test the validity of our hypothesis on this task by presenting a baseline approach, which is solely based on morphosyntactic knowledge with semantic knowledge excluded.

As for Uyghur, another major issue in TE extraction is the scarcity of resources. Specific to the issue, we collect and pre-process news data from corpora of semi-annual daily half-hour broadcast of "CCTV News" and "Xinjiang News" in Uyghur, then manually annotate with TIMEX3 tag set according to TimeML. On the basis of this human-annotated corpus, we construct the Uyghur TE dataset that consists of 4 types of TIMEX3. In Uyghur TE extraction, for the first time, Azragul et al. [5] investigated the form of simple and compound temporal words in Uyghur and proposed a rule-based approach which is mostly based on a dictionary and regular expressions. However, as rule-based approach exhibits the potential for simple TE extraction, but in a wide range of datasets that include different type of TEs, it shows relatively low recall rate due to limited rules.

In this article, we propose a TE extraction approach for Uyghur, where the extraction uses machine learning on the extensive set of features that are based on morphology, syntax, and semantics respectively. However, the work aims to apply semantic knowledge as a new promising information and analyze the effect of semantics through the development and evaluation of Uyghur TE extraction. In experimental phase, we explore the potential advantages of semantics over general features (morphology and syntax based) on this task by analyzing 28 features of 3 types, which are engineered following a systematic review of the scientific literature in TE extraction.

The paper is structured as follows: the next section describes extant works on TE extraction. A brief investigation and analysis of TE extraction in Uyghur are presented in Section 3. Feature engineering and proposed approaches are described in Section 4. Experimental results and competitive analysis of the approaches are reported in Section 5. Conclusions are drawn at last coupled with suggestions for further studies.

2. Related Work

There has been some initial works on extending TE extraction to other languages. A small parallel corpus of 95 Spanish-English dialogs has been annotated with TIMEX3 tags by a single bilingual annotator, based on the label at English side and adjusted to the Spanish (http://timexportal.wikidot.com/timex2). Also some initial works have been conducted on Chinese [6]. Besides, many systems for automatically labelling NL text have been developed following TIMEX3 standards.

HeidelTime [7] is a state of the art TE tagger, which uses a rule and pattern resources according to the TIMEX3 annotation standard, and extracts TEs with regular expression matching. In the experiment, HeidelTime achieved F1-score of 0.90 in SemEval-2013 sub-task of TE extraction. SUTime [8] is another temporal tagger for recognizing and normalizing TEs in English text. It is a deterministic rule-bases system developed for extensibility, which creates patterns over individual words to find numerical expressions, then uses patterns over words and numerical expressions to find simple TEs, and forms composite patterns over the recognized TEs. MedTime [9] is temporal information extraction system for clinical narratives, which uses hybrid approach of cascaded rule-based technique and machine learning technique. It exhibited F1-score of 0.88 in i2b2 temporal relation challenge task of TE extraction. ATT system [1] used big windows and rich syntactic and semantic feature for TempEval TEs and even segmentation and classification tasks. It uses a wide range of features like lexical, part of speech, dependency and constituency parse. It achieved F1-score of 0.85 in SemEval-2013 sub-task of TE extraction.

As is stated above, approaches related to TE extraction are mostly focused on morphosyntactic knowledge. Accordingly, those morphosyntactic features help TE extraction system gain a high performance. However, the high performance obtained is ascribed to the inclusion of word-trigger list and these pre-defined word lists that are possible to be seen in TE are very pivotal. To our knowledge, the application of word-trigger list could be become a novel form of domain-specific lexical semantics, as the application of lexical semantic resource such as semantic network has the advantage over word-triggers [10]. A common resource such as WordNet [11] takes not only the lexical semantics of a word in a specific domain (e.g., time/eventuality) but also the semantic meaning of a word within a specific domain, encoded in a lexicon with a sematic network structure. In this work, we use WordNet to build a set of named TEs, such as “Christmas Day” and “Thanksgiving Day”, as well as to expand a list of temporal triggers by adding some local Uyghur time words, based on all hyponyms of calendar_day synset.

3. Temporal Expression in Uyghur

Uyghur language is a very complex form of language which has various morphological systems, and always adopts various grammatical forms to express the whole process of event and to understand the ins and outs of events in time. Basically, a TE in Uyghur is composed of one or more words which collectively represent a point or a duration of frequency in time. Known and widely used Uyghur time words include date and time formats, names of days, months and seasons, etc. Also, words which quantify or modify time are also considered a part of a TE. Such words and phrases indicate TEs in Uyghur as follows:

- Temporal noun: (day) كۈن, (month) ئاي, (year) يىل, (hour) سائەت, (minute) مىنۇت, (second) سېكۇنت, (century) ئەسىر, (quarter) چەيسل, (week) ھەپتە, etc. Uyghur time nouns have morphological changes in person, thus present different forms in the sentences.
- Time adverb: (sometime) گاھ, (always) ھەمىشە, (from now on) ئەمدى, (a while) بىردەم, (often) ھامان, (permanent) مەڭگۈ, (usually) دائىم, etc. Uyghur time adverbs generally do not have morphological changes, but there are very few adverbs showing a less meaning of time when connected with an affix.
- Compound temporal word: (today) بۈگۈن, (this year) بۇيىل, (from tomorrow) ئەتىدىن باشلاپ, (a year from 2012 to 2014) 2012-يىلدىن 2014-يىلغىچە, (tomorrow at noon) ئەتە چۈش, (till tomorrow) ئەتىگە قەدەر, etc.

In this paper, we have two basic objectives as follows:

(1) The detection of the existing timexes in given Uyghur raw text: to determine a boundary and extent of text fragments, which are composed of one or more word units, which indicate a proper timex in the given Uyghur text. So given a document D, words w in D, it is necessary to ascertain whether *w* is in a TIMEX.

(2) Classification of the detected timexes: To classify the recognized Uyghur timexes as one temporal expression class, which is presented in the TimeML annotation standard and briefly shown in Table 1. In certain document D, there should be a mapping named I: $t \rightarrow \mathcal{X}$, where t is set as the detected timexes in D, in which $\mathcal{X} \in X$.

Table 1. Types of TE in Uyghur

Class	Example	Example (English)
DATE	2016-يىلى 3-ئاينىڭ 23 كۈنى ؛ جۈمە	March 23, 2016; Friday
TIME	ئۇچكە ئون مىنۇت ؛ سائەت بەشتە	Ten minutes to three; At five
PERIOD	2 ئاي ؛ 48 سائەت	2 months; 48 hours
FREQUENCY	ھەپتىدە ئىككى قېتىم ؛ يىلدا بىر	Twice a week; once a year

In order to deal with the two basic goals of this task, we set the delimitation or boundaries of TE and assign it a proper TIMEX3 type, so as to tag a set of words which are potentially Uyghur TE in NL text. The datasets presented in this work used brackets to delimit the set of words forming an actual TE in each sentence. Each bracketed TE holds a value indicating the type of the enclosed TE, namely TIME, DATE, DURATION, and FREQUENCY (SET). Some samples from the dataset are given to highlight the prospective result of Uyghur TE extraction. Table 2 illustrates some annotated sentences in part of Uyghur TE dataset.

Table 2. Uyghur TEs with TIMEX3 tags in sample sentences

DATE [**1988-يىل**] مەن ئاقسۇدا تۇغۇلۇپ ، DATE[**كېيىنكى يىلى**] ئۆيدىكىلەر ئونسۇغا كۆچۈپ كەپتۇ. DATE[**2 ييرىم يىلدىن كېيىن**] مىنىڭ سىگىلىم بۇ دۇنياغا كۆز ئاچتى.
DATE [**مۇشۇ دۈشەنبە**] غۇلجىدا قاتتىق يامغۇر ياغدى. شۇ سەۋەبلىك ، ھەممە ئورگانلا DURATION[**بىر ھەپتە**] خىزمەتتىن توختىدى.
بۇ قېتىمقى ئىمتىھان ۋاقتى DURATION[**سائەت 4تىن 5 ييرىمغىچە**] ئىكەن. ئۇنىڭدىن باشقا ، TIME[**6 ييرىمدا**] سىنىپ يىغىنى ئېچىلىدىكەن. ئەمدى SET[**ھەپتىدە ئىككى قېتىم**] ئېچىلىدىكەن.

The architecture of Uyghur TE extraction is summarized in Fig. 1. First, documents are preprocessed and then ready to be used for training model according to specific features given. Once the models are generated, the system uses them to annotate raw text. However, we learn the models using three approaches, Baseline (morphology only), Morphosyntax (morphology & syntax), and Semantic (morpho-syntax & semantics). Thus, experimental result difference will reflect the contribution of each approach we used on Uyghur TE extraction.

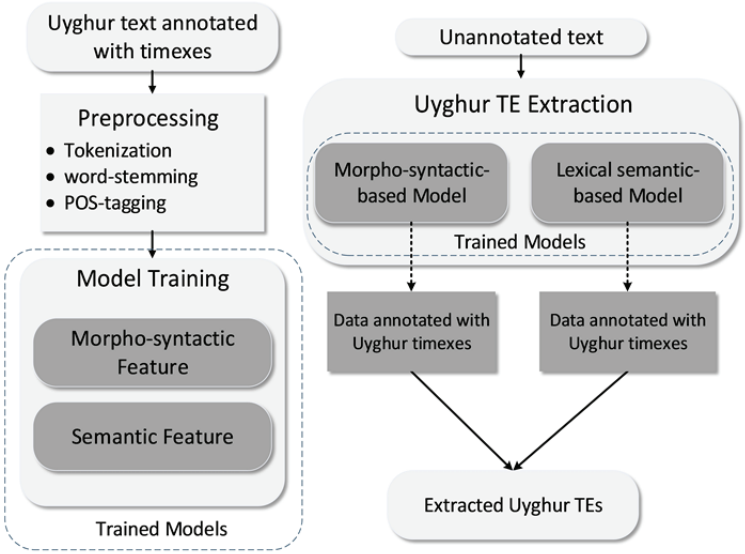

Fig. 1. Architecture of Uyghur TE extraction.

4. Uyghur Temporal Expression Extraction

4.1 Extraction Method

In TE extraction, the detection of boundary or extent of Uyghur TE in the text is a key problem to solve. In this paper, we consider the TE detection as a sequence labeling task which also can be seen as a Named Entity Recognition (NER) problem, since NER can represent a supervised sequence labeling problem [12]. For which we suppose that an input sequence of token $T_1^n = t_1 t_2 t_3 \dots t_n$, the Uyghur TE extraction is to create a label sequence $L_1^n = l_1 l_2 l_3 \dots l_n$, where l_i either belongs to the set of predefined Uyghur TE class or is not actual TE. The general label sequence l_1^n shows the highest probability of occurrence for the token sequence T_1^n between all potential label sequences. This can be written as:

$$\hat{L}_1^n = argmax\,\{Pr(L_1^n | T_1^n)\} \quad (1)$$

By virtue of chunking methodology, we use IOB2 labeling scheme [12] to tag our corpus (IOB2 represents the beginning of a TE (B), inside of a TE (I), outside of a TE (O) and sometime the E is used with the last). In this scheme, each sentence contains a word at the beginning followed by its IOB label. The label encodes the Uyghur timexes and discriminates whether the current token is inside or outside

of TE. We illustrate labeling problem by showing a sentence "*Roshen will arrive in America by October 20*" which contains some TEs in Table 3.

Table 3. Uyghur TE recognition with an IOB2 value labeling each token

Example	IOB2 value	Example (English)
روشەن	O	Roshen
10	B-TIMEX	October, 20
-	I-TIMEX	
ئاينىڭ	I-TIMEX	
20	I-TIMEX	
-	I-TIMEX	
كۈنىگىچە	I-TIMEX	
ئامېرىكىغا	O	America
يېتىپ	O	Arrive in
بارىدۇ	O	Will
.	O	.

Generally, sequence labeling task always uses machine learning technique to learn a model by observing annotated training examples. Among the supervised learning algorithms for this task, CRF performs well in a number of NLP applications, so we decide to use it for generating the model. CRF [13] is a statistical modeling tool for pattern recognition and machine leaning using structure prediction. In this model, we assume that X is an observed input data sequence to be labeled, and Y is a random variable over the corresponding label sequence. CRF model intends to find the label Y which maximizes the conditional probability $P(Y|X)$ for a token sequence x, and it can be seen as a generalization of maximum entropy and hidden Markov model that defines a conditional probability distribution taking the following form:

$$p(y|x) = \frac{1}{Z(x)} exp\left(\sum_{k=1}^{K} \lambda_k . f_k(y, x)\right) \tag{2}$$

$$Z(x) = \sum_{y \in Y} exp\left(\sum_{k=1}^{K} \lambda_k . f_k(y, x)\right) \tag{3}$$

where K is the number of features, x represents the observation sequence, y represents the label, and f_k and λ_k represent the feature function and the learned weight for each feature function, respectively.

4.2 Feature Engineering

Feature engineering is a foremost task of TE extraction for all classifiers. Moreover, the success rate in applying CRF to TE extraction principally depends on the quality of features. Regarding Uyghur language analysis level, we extract the features and classify them into general features and semantic feature. General features are most often used for TE extraction. Now, we describe the following general features used to train the model.

- **Morphological:** It includes the token, stem and POS tag in a context with at most a 5-window (-2, +2), in addition to token without letter or numbers. It achieves a good result in other NLP tasks. Furthermore, we add explicitly hand-crafted rules to match the Regex (regular expression), such as present reference, future reference, fuzzy quantifiers, modifiers, temporal adverbs and prepositions [14]. Word-segmentation, POS tagging and stemming were conducted using Modern Uyghur stemmer, MeCab-Uyghur for morphology analyzer [14].
- **Syntactic:** There are various Uyghur TEs included in particular types of phrases, such as prepositional phrase (PP) and noun phrase (NP), etc. This feature includes a token that belongs to specific one of these phrases, whose value is the key for deciding which token could be part of an Uyghur TE. This feature is extracted using Uyghur sentence constituent parser [15].

A representative semantic feature used to improve the proposed TE extraction is described as follows:

- **Lexical semantics**: A word level semantics gained form WordNet [11], which is a lexical database whose basic structure is the synset, a set of synonym words indicating an underlying lexical conception. The majority of temporal nouns included in TE are hyponyms of time, time-period (duration) or time-unit, and these time concepts are placed at the fourth level from the top concept (i.e., entity). The distribution of classes and instances over the WordNet lexical database associates with temporal categories such as TIME, DATE and DURATION or TIME PERIOD, which are the most common sense for time related concept. Many of the TEs contain words with time-related values, which will increase the probability of representing TEs for words that obtain such values, even if they do not occur in training data, for which it favors generalization to the most extent.

We, therefore, consider the lexical semantics as a feature. Table 4 illustrates some words with time-related values in WordNet.

Table 4. Uyghur time-related words in WordNet

Uyghur	English	Hypernyms hierarchy in WordNet
ئەسىر	Century	=> **time-period** => Measure => Abstraction => Entity
مىنۇت	Minute	=> **time-unit** => Measure => Abstraction => Entity
نەۋرۇز	Nowruz Festival	=> Day => **Calendar day** => time-period => Measure => Abstraction => Entity

While WordNet is one of the most semantically rich English lexical databases that is broadly used as an additional resource in many researches. Yet, still some efforts have been made in constructing multilingual WordNet [16-18]. Nonetheless, there is a limited number of languages that have successfully built their WordNets. Against this background, in this paper, we attempt to construct the lexical databases for Uyghur whose lexical conception is mainly based on temporal entities.

Uyghur is a resource-scarce language, for which we devise a time conception-based WordNet (TCBW) which only consists of temporal entity semi automatically and adapt it to the Uyghur TE extraction. Based on the Princeton WordNet (PWN) [11], we develop a simple approach to build a TCBW for Uyghur, by means of existing bilingual dictionaries and human translation. Then we automatically align all PWN's synsets which only contain temporal nouns to equivalent Uyghur synsets

through the bi-lingual dictionary. Once the synset alignment between the two languages has been finished, we can completely get synsets and relations for Uyghur TCBW. But some particular Uyghur time concepts which do not appear in PWN will be inserted according to the sense. Table 5 shows the distribution of word classes in Uyghur TCBW with respect to TIMEX3 types (namely, DATE, TIME, and DURATION), compared to the distribution of the English classes in PWN.

Table 5. Types of temporal expression in Uyghur

	#English classes	#Uyghur classes
DURATION	1,054	348
DATE	363	102
TIME	60	28

All features used in the experiment are summarized in Table 6 in detail.

Table 6. List of features used in experiments

Feature	Example
Morphological type	
Token	"بۈ" ⟶ 'بۈ'
5-window (-2, +2)	
Stop-word	"مەن", "ئۇ" ⟶ 'O', 'O'
Stem	"سەھەردە", "يازدىكى" ⟶ 'سەھەر', 'ياز'
POS-tag	"كۆز", "كەلدى" ⟶ 'N', 'V'
Suffix	"يازغۇچى" ⟶ 'غۇچى'
Ordinal number	"بىرىنجى", "ئىككىنجى", "ئۈچىنجى"
Cardinal number + period	"2 يىل", "7 قېتىم", "4 ئاي"
Contains only digits	"2016", "05", "4"
Festival expression	"روزا ھېيىت", "قۇربان ھېيىت", "نەۋرۈز بايرىمى"
Temporal future trigger	"كېيىنكى", "كىلەر", "ئالدىمىزدىكى"
Temporal fuzzy quantifier	"بىر قانچە", "تەخمىنەن", "بىر نەچچە"
Literal number	"نۆل", "بىر", "ئىككى", "ئۈچ"
Month	"فېۋرال", "مارت", "1-ياپىر"
Temporal past trigger	"بۈرۈن", "ئالدىدا", "ئىلگىرى"
Temporal period	"ئەسىر", "يىل", "ھەپتە"
Part of the day	"سەھەر", "چۈش", "كەچ"
Temporal present trigger	"ھازىر", "ئاخشام", "نۆۋەتتە"
Season	"قىش", "ئەتىياز", "ياز"
Time	"سائەت 11:50", "4:10 چۈشتىن كېيىن"
Weekday	"دۈشەنبە", "سەيشەنبە", "چارشەنبە"
Year	"1980", "ئىككى مىڭ ئون ئالتىنجى يىلى", ""
Syntactic type	
Lexical chunk	"بىر", "يىل", "قالدى" ⟶ 'B-NP', 'I-NP'
Prepositional noun phrase	"كۆزدىن", "بۇيان" ⟶ 'B-PNP', "I-PNP"
Lexical semantic type	
First sense	"ياز" ⟶ synset("summer", n)
Second sense	"ياز" ⟶ synset("write", v)
Hypernym	"ياز" ⟶ synset("summer", n) ⟶ synset("season", n)
Hyponym	"ۋاقىت" ⟶ synset("time of day", n) ⟶ synset("morning", n)

5. Experiments and Results

In this section we present the experiments performed, and particularly describe the data, evaluation metrics, and results.

5.1 Setup

Model Selection: We conduct an extensive experiment by combining 27 features mentioned above into three different models and assess if there is any statistical difference among models generated by repeating the features combination. In this way, we are allowed to select the model that outputs the highest F1-measure in Uyghur TE extraction among the three listed models.

- Model 1: Morphological only (Baseline)
- Model 2: Morphological + Syntactic
- Model 3: Morphological + Syntactic + Lexical semantics

Dataset: In Uyghur TE extraction, currently we have no standard datasets that enable our results to be compared with other experimental results. However, we use the human-annotated data of 6.74 MB, collected from corpora of semi-annual daily half-hour broadcast of "CCTV News" and "Xinjiang News" in Uyghur, as well as construct Uyghur TE dataset for this task. In Table 7, we give a brief description of our sample dataset. #Uyghur TEs stands for the actual number of temporal expressions found in the dataset.

Table 7. Types of temporal expression in Uyghur

Usage	#Docs	#Words	#Uyghur TEs
Training	400	1,322,972	3,508
Test	40	12,730	365

Evaluation Metrics: Performance of Uyghur TE extraction is evaluated based on the criteria used in TERN-2004. Two standard measures, Precision (P) and Recall (R) are used for evaluation, where P is the measure of the number of Uyghur TEs correctly identified over the number of TEs identified and R is the measure of a number of Uyghur TEs correctly identified over an actual number of Uyghur TEs. F1-measure (F) is a harmonic mean of P and R.

$$F = \frac{(\beta^2 + 1)PR}{\beta^2(P + R)}$$

5.2 Results and Analysis

Three different experimental settings have been evaluated as a combination of different features, namely Model 1, Model 2, and Model 3. Table 8 shows the results of extracted Uyghur TEs and Table 9 presents the overall performance of three different models on the proposed task.

As is shown in Table 9, for the first, the baseline model only including morphological features achieved 63.02%, 74.50% and 68.20% for Precision, Recall, and F1-measure, respectively. Although morphological information is very useful, without any post processing, the model is unable to extract

TE from the rest of the text. As the Example (1) mentioned in the introduction, the ambiguity in morphological level is a negative effect that has reduced the performance.

Table 8. Results of extracted Uyghur TEs

	#Uyghur TEs	#Correct	#Incorrect	#Missing
Model 1	357	225	55	77
Model 2	354	265	48	41
Model 3	348	305	17	26

Table 9. Performance (%) of three different models on Uyghur TE extraction

	Precision	Recall	F1-measure
Model 1	63.02	74.50	68.20
Model 2	74.85	86.60	80.30
Model 3	87.60	88.90	88.20

In the second experiment, the model including morphological and syntactic features exhibited an improved performance and obtained 74.85%, 86.60%, and 80.30% for Precision, Recall, and F1-measure, respectively, by adding syntactic parsing related feature. In this scenario, syntactic information indicates whether a word belongs to the phrase (i.e., NP, ADJP, or ADVP). This is useful for detecting more words which may be part of TE. In Example (2), this feature indicates that the double underlined word can also participate in a TE. Generally, if a NP is governed by a PP, the heading prepositions may also be essential to increasing the probability of the NP being a TE. Model 2 identifies more TEs producing high Recall by means of Uyghur sentence constituent analyzer.

Example (2)	(S (NP ئۇ) (VP ئاغرىپ ياتتى (PP بېرى (NP ئىككى يىلدىن))))
	ئۇ ئىككى يىلدىن بېرى ئاغرىپ ياتتى.

In the third experiment, the model, which is a combination of morphosyntactic and lexical semantic features, presented 87.60%, 88.90%, and 88.20% for Precision, Recall, and F1-measure, respectively, and significantly improved the performance with the highest F1-Measure as well as with a slight increase in Recall. In another way, we can count this model as an offset increasing the probability of representing TEs for words that have never seen in training data.

However, Model 3 obtained much higher results in Uyghur TE extraction. The significant improvement produced by lexical-semantic feature over baseline and syntactic feature proved our hypothesis that lexical semantics is beneficial for TE extraction. A somewhat surprising finding is that lexical semantic feature ameliorates the problem of morphosyntactic ambiguity and aids in generalization.

Regarding the errors unsolved by the proposed approaches in TE extraction, it is required to conduct a language analysis beyond semantics.

6. Conclusions

We have presented a TE extraction system in Uyghur and studied the application of semantic networks to the proposed extraction task. For this purpose, three approaches have been defined:

Morphology-based approach as a baseline; syntax-based approach using Uyghur sentence constituent analyzer; and lexical semantic-based approach using TCBW for Uyghur. The three approaches have been evaluated in the proposed extraction task. To prove the viability of our approach, we presented the Uyghur TE dataset, on which we tested TE extraction system. From the three experiment settings, the proposed approach that mostly highlighted in this work obtained 0.87 for Precision, 0.89 for Recall, and 0.88 for F1-measure and outperformed the general approaches which are based morphosyntax in Uyghur TE extraction.

The results have confirmed that exploiting the semantics to TE extraction: (1) ameliorates the performance of morphosyntactic approaches, particularly, aids in tackling morphological ambiguity and helping generalization, and (2) presents a substantial high extraction performance as compared to the other approaches.

The final results could lead us to pay attention to some potential problems of further work. On the one hand, due to the lack of local standard TimeML corpus for Uyghur, we will confront the problem of the lack of annotated dataset which directly results in the low performance in TE extraction. Hence, this study will be mostly focused on constructing more corpora by exploiting a semi-automatic processing method. On the other hand, we plan to expand our semantic feature using other kinds of semantics knowledge that have been seen very advantageous in recent studies [19]. Generating a model with more semantic features could substantially decrease the ambiguity in TE.

Acknowledgement

The work in the paper is supported by the National Nature Science Foundation of China (No. 61662081, 6186020472) and key project of National Language Commission (No. ZD1135-28); Natural Science Foundation of Xinjiang Uyghur Autonomous Region (No. 2017D01A58); National Social Science Foundation of China (No. 14AZD11); Social Science Foundation of Xinjiang Uyghur Autonomous Region (No. 2016CYY067); National Language Resource Monitoring & Research Center of Minority Languages (No. NMLR201602); Youth Sci-Tech Innovation Talents Training Project of Xinjiang (No. QN2016BS0365). The work is also supported by the key lab of network security and opinion analysis, and the key lab of data security.

References

[1] H. Jung and A. Stent, "ATT1: temporal annotation using big windows and rich syntactic and semantic features," in *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM), and the 7nth International Workshop on Semantic Evaluation (SemEval)*, Atlanta, GA, 2013, pp. 20-24.

[2] N. Chambers, "NavyTime: event and time ordering from raw text," US Naval Academy, Annapolis, MD, 2013.

[3] P. Jindal and D. Roth, "Extraction of events and temporal expressions from clinical narratives," *Journal of Biomedical Informatics*, vol. 46, pp. S13-S19, 2013.

[4] M. Filannino and G. Nenadic, "Temporal expression extraction with extensive feature type selection and a posteriori label adjustment," *Data & Knowledge Engineering*, vol. 100, pp. 19-33, 2015.

[5] Azragul, A. Murat, and Y. Abaydula, "Research on method for Uyghur temporal word recognition," *International Journal of Database Theory and Application*, vol. 9, no. 1, pp. 209-216, 2016.

[6] J. Lin, D. Cao, and C. Yuan, "Automatic TIMEX2 tagging of Chinese temporal information," *Journal of Tsinghua University*, vol. 48, no. 1, pp. 117-120, 2008.

[7] J. Strotgen and M. Gertz, "HeidelTime: high quality rule-based extraction and normalization of temporal expressions," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, Uppsala, Sweden, 2010, pp. 321-324.

[8] A. X. Chang and C. D. Manning, "SUTime: a library for recognizing and normalizing time expressions," in *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey, 2012, pp. 3735-3740.

[9] Y. K. Lin, H. Chen, and R. A. Brown, "MedTime: a temporal information extraction system for clinical narratives," *Journal of Biomedical Informatics*, vol. 46, pp. S20-S28, 2013.

[10] H. Llorens, E. Saquete, and B. Navarro-Colorado, "Applying semantic knowledge to the automatic processing of temporal expressions and events in natural language," *Information Processing & Management*, vol. 49, no. 1, pp. 179-197, 2013.

[11] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[12] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3-26, 2007.

[13] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning*, Williamstown, MA, 2001, pp. 282-289.

[14] A. Abdurehim, "Automatic inference of affix variants in Uyghur based on POS-tagging corpus," *Computer Knowledge and Technology*, vol. 12, no. 28, pp. 171-173, 2016.

[15] Nurehmet, Azragul, and Y. Abaidulla, "The research of modern Uyghur language sentence constituents analysis technology," *Computer Engineering and Science*, vol. 2015, no. 12, pp. 2339-2344, 2015.

[16] H. Isahara, F. Bond, K. Uchimoto, M. Utiyama, and K. Kanzaki, "Development of the Japanese WordNet," in *Proceedings of the International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 2008.

[17] M. Montazery and H. Faili, "Automatic Persian wordnet construction," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, Beijing, China, 2010, pp. 846-850.

[18] S. Thoongsup, K. Robkop, C. Mokarat, T. Sinthurahat, T. Charoenporn, V. Sornlertlamvanich, and H. Isahara, "Thai WordNet construction," in *Proceedings of the 7th Workshop on Asian Language Resources*, Singapore, 2009, pp. 139-144.

[19] O. Kolomiyets, S. Bethard, and M. F. Moens, "Model-portability experiments for textual temporal analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers*, Portland, OR, 2011, pp. 271-276.

Alim Murat https://orcid.org/0000-0001-8510-7808

He received B.E. and M.S. degrees in School of Computer Science and Technology from Xinjiang Normal University in 2011 and 2014, respectively. He finished his Ph.D. degree in Computer Science from Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science in June 2017. Since July 2017, he has been working in the Xinjiang Normal University as a lecture. His current research focus includes natural language processing and semantic web.

Azharjan Yusup https://orcid.org/0000-0002-5641-2229

He received his B.S. degree in Computer Science from Xinjiang Normal University in 2012 and was enrolled by the Xinjiang Normal University as a master student in Computational linguistics in July 2017.

Zulkar Iskandar https://orcid.org/0000-0002-3988-5817

He received his B.S. degree in Computer Science from Xinjiang Agricultural University in 2012 and was enrolled by the Xinjiang Normal University as a master student in Computational linguistics in July 2017.

Azragul Yusup https://orcid.org/0000-0002-8264-1033

She received her Ph.D. degree in Computer Science from Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science in 2016. Since July 2016, she has been working in the Xinjiang Normal University as a lecture. Her current research focus includes natural language processing and computational linguistics.

Yusup Abaydulla https://orcid.org/0000-0001-5015-2481

He is a Prof. and director of the key lab for network security and sentiment analysis. He is also a lead researcher in the School of Computer Science, Xinjiang Normal University. His current research focus includes natural language processing and computational linguistics. He has published more than 50 papers in a various journals and refereed conferences.