Time Phrase Parsing for Chinese Text with HowNet Temporal Information Structure

Hong-mei Ma, Xiao-yun Wang and Li Qin
Library, China Defense Science and Technology Information Center, Beijing, China

Keywords: Information Retrieval, Temporal Information, Time Phrase Parse, Chinese Text.

Abstract: Time phrase parsing is useful to analyze Chinese text, because it is usually omitted expression in the Chinese text for the coherence and the cohesion. It is important to gain the temporal information, then it can be used to generate the tense of the verb. In this paper, with temporal information structures of HowNet, time phrases are divided into three categories, and a structure is designed to represent the temporal information of time phrases. Then, it puts forward a method to parse time phrases in Chinese text.

1 INTRODUCTION

As we all know, there isn't morphologic changing in Chinese. Temporal informal implies in the Chinese text. Since temporal system is objective to every kind of natural language, we could gain the temporal information from the Chinese text. It has been proved that the temporal system has three components, phase, tense and aspect (Zhou 2000). Phase is a property of the event described by the semantic of the verb, tense is a temporal relation between the event time (ET) and the speech time (ST) or another reference time (RT), and aspect is the state of the event, which depends on the point-of-view of the observation (M & G 1996). Temporal relations are precedence, simultaneous and include-in. Note that the tense discussed in the paper is different from the syntactic category tense in English, values of it are happen-ed, happen-ing and happen-will. Happen-ed represents ET is precedence to ST (or RT), happen-ing means ET is simultaneous to ST (or RT), or ET is included in RT, happen-will indicates ST (or RT) is precedence to ET. In the Chinese language time phrase is a syntactic approach to represent the tense of the verb (Gong 1995), it often has temporal references among the sentences in the Chinese text for the coherence and the cohesion, so it is necessary to parse the time phrases in the Chinese text in order to determine the tenses of the verbs in the text.

HowNet is a bilingual general knowledge-base describing relations between concepts and relations between the attributes of concepts. It has more than 2,000 sememes to define the concepts and proposes information structures to reveal the meaning of phrases at the semantic level (Dong 2007). Chinese information structures database of HowNet (short for CISD) is derived from large scale natural language corpus, which gives 268 information structures and more than 10,000 examples in the latest version (Dong 2010). In CISD there are 23 temporal information structures and 1,029 examples of time phrases, then we analyse these temporal information structures and experiment with the examples. In order to computing semantic of time phrases, we category time phrases and design a semantic structure, named TPSS, to represent the temporal information of time phrases.

In the paper, time phrases are categorized into three in section 2, then, a method is put forward to parsing time phrases in section 3. Finally, we implement a system to parsing time phrases in Chinese texts.

2 TIME PHRASE CATEGORIZATION

Constituents of time phrases include time noun words, for example, 昨天(zuo2 tian1, yesterday), time expression that composed of number and time-quantifier, such as 2013 年 5 月(er4 ling2 yi1 san1 nian2 wu3 yue4, May 2013), in which 年

Ma H., Wang X. and Qin L..
Time Phrase Parsing for Chinese Text with HowNet Temporal Information Structure.
DOI: 10.5220/0005094902940299
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2014), pages 294-299
ISBN: 978-989-758-048-2

(nian2, year) and 月(yue4, month) are two time-quantifiers, time order words, for example, 前(qian2, preceding), and part-of time words, for instance, 初(chu1, beginning) in Chinese. Semantic of time phrases is illustrated by temporal information structures of HowNet (short for TIS) intensively, for example:

(1) 昨天-上午(zuo2 tian1 shang4 wu3, yesterday morning)
TIS=(时间|time,特|special,日|day) [whole] ← (时间|time, 特|special, 晨/午/晚|morning/afternoon/night)

(2) 会-前(hui4 qian2, before the meeting)
TIS=(事情|things) ← [time] (时间|time)

In (1) it describes the ***morning*** in ***yesterday***, concept ***yesterday*** is the *whole* of concept ***morning***. Example (2) is about the *time* ***before*** the time when the ***meeting*** is began. According to the meaning of time phrases, we category temporal information structures of HowNet and time phrases into three kinds, date, presudodate and time interval.

2.1 Date

Date is the numbers or words used to talk about a particular day, month, and year. It has the definite value on the axis of time. The formation of the time phrase expressing the date in Chinese is shown as follows:

Date::= [<Time-unit>]* (1)
Time-unit::= <time-num><time-quantifier> (2)

The time-quantifiers of the dates include 世纪(shi4 ji4, century), 年(nian2, year), 季(ji4, season), 月(yue4, month), 旬(xun2, ten days), 周(zhou1, week), 日(ri4, day), 日段(ri4 duan4, part of a day), 时(shi2, hour), 分钟(fen1 zhong1, minute) and 秒钟(miao3 zhong1, second) (Dong 1999). The time-unit that composed of the former time-quantifier is larger than the time-unit composed of the latter time-quantifier. Moreover, 季 has 4 values: 春(chun1, spring), 夏(xia4, summer), 秋(qiu1, autumn) and 冬(dong1, winter), which have the precedence relation. And 日段 also has 4 values: 晨(chen2, morning), 午间(wu3 jian1, noon), 下午(xia4 wu3, afternoon) and 晚(wan3, night), which also have precedence relation. Finally, every time-unit of the dates could be divided into three parts again, which are the beginning part, the middle part and the ending part, and there is precedence relation between these parts.

The dates have two properties in the Chinese text. One is that the dates often only contain the larger time-unit when it isn't necessary to point out the detail time. The other is that the dates often have the temporal references in the Chinese text. The larger time-unit of the date would be omitted when it is same as the larger time-unit of the date in the above sentences.

2.2 Presudodate

Presudodate is composed of a based-date and an offset, in which the based-date is a date, and the offset consists of direction and size. Because the offset and the based-date of the presudodate are often indefinite, the presudodate hasn't definite value on the axis of time. However, the presudodate could be transformed to the date when the base-date and the offset are given. Presudodate is defined as follows:

Presudodate ::= <Based-date><Offset> (3)
Offset ::= <Offset-dir><Offset-size> (4)
Date=Presudodate.Based-date + Presudodate.Offset (5)

Expression (5) shows the process of transforming the presudodate to the date. When the direction of the offset is same as the direction of the axis of time (the flow of time), the size of the offset is positive. Well, the size of the offset will be negative when the direction of the offset is opposite to the direction of the axis of time. And when the size of the offset is zero, it implies that the presudodate denotes the based-date, or the larger time-unit of the presudodate and that of the based-date are same, for example, *今天*(jin1 tian1, today), *今天下午*(jin1 tian1 xia4 wu3, this afternoon).

Time phrases expressing the presudodates have 4 forms generally in the Chinese text:

(1) Have the based-date and the offset direction. For example, 2008 *年6月1日前*(er4 ling2 ling2 ba1 nian2 liu4 yue4 yi1 ri4 qian2, before June 1st 2008). This kind of presudodate could be transformed to the date with formula (5) after receiving the size of the offset.

(2) Have the offset including the direction and the size, and take the latest date in the above sentences as its based-date. For example, *"1999年，…。两年后，…。"*(yi1 jiu3 jiu3 jiu3 nian2, …liang2 nian2 hou4,…，… in 1999. … two years later …).

(3) Have the offset, and take the date that the event happened as its based-date. For example, *"第二次世界大战以后"*(di4 er4 ci1 shi4 jie4 da4

zhan4 yi2 hou4，after The Second World War).

(4) Have the offset, and take the date that the text was written as its based-date. For example, *"明天"*(ming2 tian1, tomorrow), its offset direction is positive and its offset size is one day.

2.3 Time Interval

What the third kind of time phrases represents is a time interval, which has two different forms in the Chinese text. One is composed of the beginning date and the ending date, such as *80 年代末 90 年代初* (ba1 shi2 nian2 dai4 mo4 jiu3 shi4 nian2 dai4 chu1, from the ending of 1980s to the beginning of 1990s), the other just represents the size of the time interval, for instance, *十天*(shi2 tian1, ten days). By the way, the size of the offset of the presudodate mentioned above is a time interval.

3 TIME PHRASE PARSING

3.1 TPSS

TPSS is the abbreviation for the Time Phrase Semantic Representation Structure. It contains the temporal information of a time phrase, its definition is:

$$\text{TPSS} ::= \langle \text{Semcate, Semcontent} \rangle \quad (6)$$

$$\text{Semcate} \in \{\textit{date, presudodate-basedate, presudodate-timeinterval}\}$$

$$\text{Semcontent} ::= \langle \text{Date, Offset-dir, Offset-size} \rangle \quad (7)$$

$$\text{Offset-dir} ::= \textit{nega} \mid \textit{zero} \mid \textit{posi} \quad (8)$$

$$\text{Offset-size} ::= [\langle \text{offset-size-unit} \rangle]^* \mid \langle \text{how-adv} \rangle \quad (9)$$

$$\text{offset-size-unit} ::= \langle \text{time-num} \rangle [\langle \text{special-quantifier} \rangle] \langle \text{time-quantifier} \rangle \quad (10)$$

Semcate is the category of the time phrase that we have discussed in section 2. Semcontent is the temporal information of the time phrase defined as a triple including date, offset-dir and offset-size. How-adv is the adverb representing the degree of the quantity, for instance, 左右(zuo3 you4, or so), 整 (zheng3, exactly). And the special-quantifier is some quantifier that could be used between the number and the time-quantifier.

TPSS is an enclosure of the temporal information of the date, the presudodate and the time interval. When the Semcate of TPSS is *date*, the offset-dir and the offset-size is zero. When TPSS represents a presudodate with the based-date, date of the Semcontent is the based-date of the presudodate. And when the Semcate of TPSS is *presudodate-timeinterval*, there would be two dates in TPSS, one is the beginning date, the other is the ending date, and the size of the time interval would be represented by the offset-size.

3.2 Mliurn

Linguistic context of Chinese text is important to the Chinese text understanding. HowNet gives the static general knowledge about the world, but the dynamic context knowledge about the text should also be collected in the process of text analysis. In the paper, a structure named MLIURN (Multilevel Information Unit Relation Network) is designed to represent the context knowledge of the Chinese text, see Figure 1:

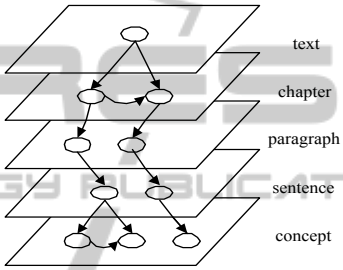

Figure 1: Multilevel Information Unit Relation Network.

As indicate in Figure 1, MLIURN has 5 levels, every level has information units and the relations. On one hand, information unit at the low level is part of information unit at the high level. On the other hand, information units and relations at different level have different contents that representing the knowledge of natural language from different level. For example, at the concept level, information unit is one concept which has syntactic properties, semantic properties and the number of the sentence in which the concept is contained. The relation between two information units is the semantic relation between the two concepts. Concepts expressed in the concept level include event concepts, entity concepts and time concepts. TPSS is one information unit at the concept level. Well, information unit at the text level describes the information of the text, such as the text structure, the text category and the theme of the text.

3.3 Time Phrase Parser

Time phrase parser is a Finite-State Machine including a table of input symbol sets and a function of state transition. Every state of the parser is an object with special semantic action. First, it writes the temporal information of the time phrase into TPSS. Second, it would to search MLIURN to receive the

temporal information from the above paragraphs when there are time references in the Chinese text. The structure of the parser is showed in Figure 2.

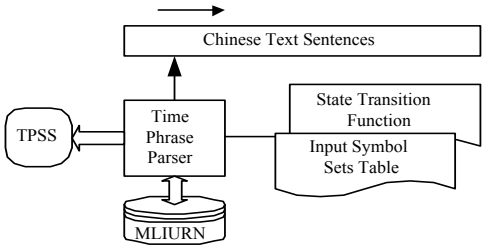

Figure 2: Chinese Text Time Phrase Parser.

We define 18 kinds of input symbol sets as indicated in Table 1.

Table 1: Input Symbol Sets Table.

A	Chinese number set i.e. “五”(wu3, five)	J	Month set i.e. “三月 ”(san1 yue4, March)
B	Digit set i.e. “8”	K	Presudodate time word set i.e. “昨天 ”(zuo2 tian1, yesterday)
C	Time-quantifier set i.e. “ 月 ”(yue4, month)	L	Presudodate prefix set i.e. “ 过去 ”(guo3 qu4, bypast)
D	Time pronoun set i.e. “那”(na4, that)	M	Presudodate suffix set i.e. “以前”(yi3 qian4, before)
E	PartofDay set i.e “上午”(shang2 wu3, in the morning)	N	Special quantifier set i.e. “个”(ge4)
F	Year Prefix set i.e. “公元”(gong1 yan2)	O	Time order word set i.e. “第”(di4)
G	Quantity adv set i.e. “左右”(zuo3 you4, or so)	P	Other time word set i.e. “国庆节”(guo2 qing4 jie2, National Day)
H	Part-of time word set i.e. “ 初 ”(chu1, beginning)	Q	Education place set i.e. “ 大学 ”(da4 xue2, university)
I	Weekday set i.e. “星期一”(xing1 qi1 yi1, Monday)	R	Verb set i.e. “毕业 ”(bi4 ye4, graduate)

Time phrase parsing algorithm is as follows:

Time Phrase Parsing Algorithm（TPPA）

Let S_i be the current state of the parser, X be one set of Table 1, and F be the state transit function.

```
1)  Input word w.
2)  Evaluate the word w whether or not
    belong to one set of Table 1.
3)  If not, goto 9).
4)  If w ∈ X, calculate the state S' to
    which Si transit by F(Si, w).
5)    If S'=NULL, return to the last
    termination state, goto 9).
6)  If S'≠NULL, set S' to Si .
7)  Perform the semantic action of Si .
8)  If Si is a termination state, write the
    result of 7) to TPSS, goto 1).
9)  If TPSS=NULL, return;
10) If TPSS≠NULL, test TPSS whether or not
    having temporal reference to the above
    paragraph.
11) If there is temporal reference, search
    MLIURN, extend TPSS with the temporal
    information of the time phrase in the
    above paragraph.
12) Write TPSS into MLIURN, return.
```

3.4 Tense Calculus

Now the tense of the verb that restricted by the time phrase could be calculated. First, some simple operators are defined, and then an algorithm (Chinese Text Verb Tense Parsing Algorithm) to calculate the tenses of the verbs in Chinese text.

DEFINITION 3.1 Date plus operator +

Let X be a set of dates, and Y be a set of offsets, date plus operator + is a transformation from (X, Y) to X .

Suppose $\forall d \in X$, $\forall o \in Y$,

$d = \{\text{time-unit}_1, \text{time-unit}_2, \ldots, \text{time-unit}_n\}$,

$o = \{\text{offset-size-unit}_1, \text{offset-size-unit}_2, \ldots, \text{offset-size-unit}_m\}$, m, n $\in$Z.

$d + o =$

$d.\text{time-unit}_i.\text{time-num} + o.\text{offset-size-unit}_j.\text{time-num}$ iff

$d.\text{time-unit}_i.\text{time-quantifier} = o.\text{offset-size-unit}_j.\text{time-quantifier}$,

$i, j \in \{1,2,\ldots,n\}$

DEFINITION 3.2 Date minus operator -

Let X be a set of dates, and Y be a set of offsets, date minus operator - is a transformation from (X, Y) to X .

Suppose $\forall d \in X$, $\forall o \in Y$,

$d = \{\text{time-unit}_1, \text{time-unit}_2, \ldots, \text{time-unit}_n\}$,

$o = \{\text{offset-size-unit}_1, \text{offset-size-unit}_2, \ldots, \text{offset-size-unit}_m\}$, m, n $\in$Z.

$d - o =$

$d.\text{time-unit}_i.\text{time-num} - o.\text{offset-size-unit}_j.\text{time-num}$ iff

$d.\text{time-unit}_i.\text{time-quantifier} = o.\text{offset-size-unit}_j.\text{time-quantifier}$, $i, j \in \{1,2,\ldots,n\}$

DEFINITION 3.3 Date compare operator R

Let X be a set of dates, and Y be a set of relations between dates, Y={*pre*, *equ*, *aft*}, date compare operator R is a transformation from (X, X) to Y .

Suppose $\forall\, d_1, d_2 \in X$,

$d_1 = \{\text{time-unit}_1, \text{time-unit}_2, \ldots, \text{time-unit}_n\}$,

$d_2 = \{\text{time-unit}_1, \text{time-unit}_2, \ldots, \text{time-unit}_n\}$,

time-unit_i is larger than time-unit_{i+1}, $i=1,2,\ldots n-1$.

i) If $d_1.\text{time-unit}_i.\text{time-num} < d_2.\text{time-unit}_i.\text{time-num}$, $R(d_1, d_2) = \textit{pre}$.

ii) If $d_1.\text{time-unit}_i.\text{time-num} > d_2.\text{time-unit}_i.\text{time-num}$, $R(d_1, d_2) = \textit{aft}$.

iii) If $d_1.\text{time-unit}_i.\text{time-num} = d_2.\text{time-unit}_i.\text{time-num}$, $i = i+1$; continue i), ii), iii).

iv) For $i = 1,\ldots n$, if $d_1.\text{time-unit}_i.\text{time-num} = d_2.\text{time-unit}_i.\text{time-num}$, $R(d_1, d_2) = \textit{equ}$.

DEFINITION 3.4 Datelization operator D

Let X be a set of dates, and Y be a set of presudodates, F be a set of the directions of the offset, F={*posi*, *nega*, *zero*}, datelization operator D is a transformation from X to Y .

Suppose $\forall\, d \in X$,

i) if d.offsetsir = *posi* , D(d) = d.basedate + d.offsetsize,

ii) if d.offsetsir = *nega* , D(d) = d.basedate – d.offsetsize.

In i), ii), d.basedate is the based-date of the presudodate, d.offsetdir is the direction of the offset of the presudodate, and d.offsetsize is the size of the offset of the presudodate.

DEFINITION 3.5 Offset direction operator Θ

Let X be a set of dates, and F be a set of the directions of the offset, F={*posi*, *nega*, *zero*}.

Suppose $\forall\, d_1, d_2 \in X$,

i) if $R(d_1, d_2) = \textit{pre}$, $\Theta(d_1, d_2) = \textit{nega}$,

ii) if $R(d_1, d_2) = \textit{equ}$, $\Theta(d_1, d_2) = \textit{zero}$,

iii) if $R(d_1, d_2) = \textit{aft}$, $\Theta(d_1, d_2) = \textit{posi}$.

$\Theta(d_1, d_2)$ indicates the direction of the offset that d_1 relative to d_2.

DEFINITION 3.6 Tense operator T

Let X be a set of dates, and S be a set of the tenses, S={ *happen-ed*, *happen-ing*, *happen-will* }.

Suppose $\forall\, d_1, d \in X$,

i) if $\Theta(d_1, d) = \textit{nega}$, $T(d_1, d) = \textit{happen-ed}$,

ii) if $\Theta(d_1, d) = \textit{zero}$, $T(d_1, d) = \textit{happen-ed}$,

iii) if $\Theta(d_1, d) = \textit{posi}$, $T(d_1, d) = \textit{happen-ed}$,

In i), ii), iii), d is the date when the text is written, so $T(d_1, d)$ indicates the tense of d_1 which relative to d.

The algorithm is defined as follows:

Chinese Text Verb Tense Parsing Algorithm

Let d be the date when the text is written, S be the current sentence, and V be the verb in S.

Step1 If S has time phrase, receive its semantic structure named $TPSS_i$ by TPPA.

(1) If $TPSS_i$ is a date, denoted d_i, compute $\Theta(d_i, d)$, and then compute $T(d_i, d)$.

(2) If $TPSS_i$ is a presudodate, denoted pd_i,

i) If pd_i has based-date,

a. If pd_i.based-date=d, infer the tense of pd_i from pd_i.offset-dir.

b. If pd_i.based-date≠d, compute $D(pd_i) = d_j$, then compute $\Theta(d_j, d)$, finally compute $T(d_j, d)$.

ii) If pd_i has offset, and take the last date in the above paragraph as its based-date, get the date from MLIURN, then perform as i).

iii) If pd_i take the date that the event happened which in the above paragraph as its based-date, get the date from MLIURN, then perform as i).

(3) If $TPSS_i$ is a time interval,

i) If $TPSS_i$ has the beginning date d_m and the ending date d_n, compute $\Theta(d_m, d)$ and $\Theta(d_n, d)$. If $\Theta(d_m, d) = \textit{posi}$, the tense is *happen-will*, if $\Theta(d_n, d) = \textit{nega}$, the tense is *happen-ed*, otherwise, the tense is *happen-ing*.

ii) If $TPSS_i$ is a time interval with the size of the interval, take the tense of the last sentence in the above paragraph as its tense.

Step2 If V is restricted by $TPSS_i$, set the tense of $TPSS_i$ to the tense of V.

Step3 If S has no time phrase, take the tense of the last sentence in the above paragraph as the tense of V.

4 EXPERIMENTS

We have built a text parsing model and a text context knowledge representation structure MLIURN in order to solve some problems of the sentence parsing model. With the parser, it has parsed the 23 temporal information structures of HowNet and the examples, analyzed time phrases in the Chinese text, and evaluated the tenses of the verbs.

The following text is taken from the newspaper BEIJING YOUTH DAILY (16 April 2014), and the analysis result is showed in Table 2.

李嘉诚“七连抛”躲避资产风波

本报讯 近日，李嘉诚旗下的和记黄埔有限公司得到伦敦市长的许可，获批在伦敦东南部的德特福德兴建房地产项目。而本月初，李嘉诚次子李泽楷以 9.28 亿美元出售其掌控的盈大地产旗下北京盈科中心。至此，从去年 8 月开始，李家在内地和香港已上演“七连抛”，“一抛一入”之间引发了市场的相关猜测。据统计，2010 年至今李嘉诚大举抛售香港和内地资产，其市值已超过总身价七成。

摘摘自《北京青年报》〔20140416〕

Table 2: Parsing Result.

Event	Time phrase	Temporal reference	Tense*
得到(de2 dao4)	近日(jin4 ri4)		happen-ed
获批(huo4 pi1)		Yes	happen-ed
兴建 (xing1 jian4)		Yes	happen-ed
出售 (chu1 shou4)	本月初 (ben3 yue4 chu1)		happen-ed
上演 (shang4 yan3)	从去年 8 月开始 (cong2 qu4 nian2 ba1 yue4 kai1 shi3)		happen-ed
引发(yin3 fa1)		Yes	happen-ed
抛售 (pao1 shou4)	2010 年至今(er4 ling2 yi1 ling2 nian2 zhi4 jin1)		happen-ed
超过 (chao1 guo4)		Yes	happen-ed

*Tense is one element of temporal system, which isn't the syntactic category tense in English. The value indicates the temporal information of the events in the Chinese text.

5 CONCLUSION

In this paper, temporal information structures of HowNet and time phrases are divided into 3 kinds, date, presudodate and time interval. Then, a structure, called TPSS, is designed to represent the temporal information of time phrases. Finally, a parser is built to parse time phrases and to infer the tenses of the verbs in Chinese texts.

Although the parsing algorithm of time phrases in Chinese text has shown satisfactory result, some difficulties are still remained, such as ambiguous word segmentation. Because of lacking for commonsense knowledge, tenses of some presudodates couldn't be calculated. We will do more work about the Chinese temporal system in the future.

ACKNOWLEDGEMENTS

We gain many ideas in the discussion with Mr. Dong Zhendong. This work would not be possible if without his advice. We would like to thank him.

REFERENCES

Qianyan Gong, 1995. Phase Tense and Aspect in Chinese, The Commercial Press.

Gagnon M., and Lapalme G., 1996. From Conceptual Time to Linguistic Time, Association for Computational Linguistics.

Chongli Zhou, 2000. Research of Natural Language Logic, Beijing University Press.

Zhendong Dong, 2007. Theoretical findings of HowNet, Journal of Chinese Information Processing, vol.21, no.4, 3-9.

Qiang Dong, 2010. A HowNet-Based Disambiguator for Chinese Syntactic Structures, Journal of Chinese Information Processing, vol.24, no.1, 60-64.

HowNet. http://www.keenage.com.