A Rule System for Chinese Time Entity Recognition by Comprehensive Linguistic Study

Hongzhi Xu
The Department of CBS
The Hong Kong Polytechnic University
hongz.xu@gmail.com

Chu-Ren Huang
Faculty of Humanities
The Hong Kong Polytechnic University
churenhuang@gmail.com

Abstract

Chinese time entity is quite complex. In this paper, we give a comprehensive linguistic study on it. Based on the analysis, we present a rule system which only considers the inner structure of Chinese time entities for the recognition. Experiments on Sinica and TempEval-2 corpus show that the rule system performs much better than the CRFs model. When using the rules as features within a CRFs model, the performance could be further improved.

1 Introduction

In SemEval-2010 competition, there is a sub task for temporal entity identification, which includes a Chinese corpus. The final goal of the task is to associate a temporal expression to a certain event. It is very important to extract all the elements for events in that it will be useful for event tracking. By identifying the time information of events will enable us to make inference on the temporal relation of different events.

In this paper, we will make a comprehensive study on Chinese time entities from a linguistic perspective and then present a rule system for recognizing them. Chinese temporal entity is very complex due to the flexible grammar of Chinese and the existence of many different time systems, such as Gregorian system, the Chinese lunar system, the Chinese tian-gan & di-zhi (GZ) time system.

Based on our linguistic analysis, we formalize a set of temporal elements that are the blocks used to construct time entities, such as century, year, month, day, hour etc. We then build a rule system that actually describe the topology of the temporal elements. For example, year follows century; month follows year. So, the model of our system is a directed graph, while a valid temporal expression is a path from one certain node to another node. The longer the path is, the more confident the recognition will be.

CTEMP (Wu et al., 2005) also used linguistic rules for Chinese temporal entity recognition. However, the focus of this work differs from them in that we aims to identify Chinese time entities which could be described with a limited set of rules and can be easily translated into a structured format, such as TIMEX3(Pustejovsky et al., 2010) standard. For this part, the set of rules in this work are more comprehensive than (Wu et al., 2005). However, we don’t include events that are used as time entities, since events intrinsically are not time entities. According to the Generative Lexicon Theory (Pustejovsky, 1995), this is a case of type coercion.

In Section 2, we will give a linguistic study on Chinese time entity expressions. In Section 3, we will construct a rule system which is mainly based on our linguistic study. In Section 4, we test rule system on Sinica and TempEval-2 corpora and give a discussion on the experimental result. Section 5 is the conclusion.

2 Chinese time entity: A linguistic study

We refer to Y.R. Chao’s book (Chao, 1968) as a starting point of our study. In China, there are different time systems, including the lunar system, TianGan-DiZhi (GZ) system, etc. In ancient China, people used the emperor’s reign to count time. When a new emperor appeared, a new period would then started.

In another perspective, people try to divide the time axis by different levels of granularity. Roughly, the whole axis can be divided into three periods: guo-qu (past), xian-zai (present) and jiang-lai (future). Smaller granularity includes century (shi-ji), year (nian), season (ji-jie), month (yue), day (ri), hour (shi), minute (fen), second (miao). Week (zhou) is a granularity that is independence-
t to year, season and month. In China, there are also jie-qi (JQ) that divides one year into 24 different periods. One month can also be divided into 3 periods (XUN); the first ten days (shang-xun), the second ten days (zhong-xun) and the left days (xia-xun). One day can also be divided into different vague phases (DP), e.g. before dawn (ling-chen), early morning (zao-shang), morning (shang-wu), noon (zhong-wu), afternoon (xia-wu), evening and night (wan-shang), midnight (wu-ye).

To compile rules for the automatic recognition of Chinese time entities, one important issue is to find out the construction regularity for each temporal element and the relations among the elements, which is also the inner structure of Chinese time entities.

2.1 Gregorian system and Chinese lunar system

Gregorian system starts from the year of Christ’s birth. Before this year, B.C. (gong-yuan-qian) is used with a number to denote time on the time axis. After this year, A.D. (gong-yuan) is used, which is also the default value. Chinese supports this system. For example, 2013-08-08 09:01:01 is said in Chinese (gong yuan) er-ling-yi-san-nian ba-yue ba-ri jiu-dian ling-yi-fen ling-yi-miao.

One hour can also be divided into four quarters (ke). However, only yi-ke (fifteen) and san-ke (forty five) are valid expressions. For the half of an hour, ban (half) is used. zheng (right) will be used as the right start of an hour. So, zheng, yi-ke, ban, san-ke are the four possible values for the KE element.

One year can be divided into four quarters (ji-du:JD) or (ji-jie:season). An ordinal number is used to refer to a certain JD, such as di-yi ji-du (the first quarter). The ordinal marker di could be omitted. So, yi ji-du is also a valid expression. Each season has its own name: spring (chun-ji), summer (xia-ji), autumn (qiu-ji) and winter (dong-ji).

For hours, day phases (DP) could be added before them. The DP is usually placed before hour, such as ling-chen san-dian (3:00am), wu-ye shier-dian (0:00). However, the boundaries of different phases are not clear, such as xia-wu/wan-shang liu-dian (6:00 in the afternoon/evening).

Century (shi-ji) can be followed by decade (nian-dai), such as er-shi-shi-ji jiu-shi-nian-dai (the 90s of 20th century). The first decade is usually called ling-ling-nian-dai (00s) or tou-shi-nian (first ten years).

If gong-yuan (A.D.) or gong-yuan-qian (B.C.) is used before century or year, then the numbers will be written as the pronunciation of the number rather than a sequence of digits. For example, gong-yuan liang-qian-ling-yi-shi-san nian is similar to be said as two thousand and thirteenth years A.D. in English. Otherwise, year 2013 will be written as er-ling-yi-san-nian (two-zero-one-three year).

Chinese lunar time system uses a similar way to denote time as the Gregorian system. However, it refers to the movement of the moon to count months. So the start of one year in lunar system is different from the Gregorian system. We can use a flag ‘&’ (nong-li) to denote the lunar system, such as & 2013-08-08. In addition, the lunar system uses chu before the day number for the first ten days of a month in order to make up of two syllables, while the day marker ri is usually omitted. For example, Aug. 8th is said ba-yue chu-ba, Aug. 11th is said ba-yue shi-yi. The lunar label nong-li can also be placed before the subsequence of year-month-day, such as nong-li wu-yue chu-wu (& 05-05), nong-li chu-wu (& 05)’ etc.

2.2 TianGan-DiZhi system

This system was invented in Ancient China based on the Chinese traditional philosophical theory. There are ten heavenly stems (tian gan: TG): jia, yi, bing, ding, wu, ji, geng, xin, ren, gui and twelve mundane branches (di zhi: DZ): zi, chou, yin, mao, chen, si, wu, wei, shen, you, xu, hai. Then, one year is denoted by a combination of two different elements circularly, which generates sixty different denotations. If we use a sequence number to denote the two elements, i.e. \( TG_{i} \) and \( DZ_{i} \), then the ith year of a circulation is defined as \( y_i = TG_i \%10 DZ_i \%12 \), where \( 0 \leq i < 60 \) and \( \% \) is the mod operation. For example, gui-si-nian (2013) can be formally denoted as year \( TG_9 DZ_5 \), or simply \( GZ_{95} \). Similarly, month, day and the Chinese hour can also be denoted like this.

The twelve DZ items are also associated with twelve animals (sheng xiao: SX): shu (mouse), niu (cattle), hu (tiger), tu (rabbit), long (dragon), she (snake), ma (horse), yang (sheep), hou (monkey), ji (chick), gou (dog), zhu (pig). So, one year can also be simplified as {animal} nian. For example,
year 2013 can be also called as *she-nian (year of snake), or formally denoted as SX5. However, this kind of expression can only be said alone. It can rarely be said with month and day, such as *she-nian wu-yue (the 5th month of year of snake).

2.3 Jie-Qi

As we have mentioned, there are also twenty four Jie-Qi (JQ) within one year: li-chun, yu-shui, jing-zhe, chun-fen, qing-ming, gu-yu, li-xia, xiao-man, mang-zhong, xia-zhi, xiao-shu, da-shu, li-qi, chu-shu, bai-lu, qiu-fen, han-lu, shuang-jiang, li-dong, xiao-xue, da-xue, dong-zhi, xiao-han, da-han. Every six JQs corresponds to and divide one season. The JQs are actually time words and included in Chinese dictionaries. JQ usually follows year element, such as er-ling-yi-san-nian qiu-fen (qiufen of 2013).

2.4 Regnal year system

Ancient Chinese people have seen a new emperor as a starting point of a new period. A number is used to count the following years after that year. The first year is called yuan-nian, the second year is called er-nian (2nd year), etc. For example, QianLong yuan-nian stands for the year when QianLong became the emperor. However, there are hundreds of emperors in the history of China, and many of them are not recorded at all. So, the list of emperors is hard to be complete. Usually, the most used regnal years refer to the Qing Dynasty.

2.5 Weekdays

Weekdays (xing-qi) are expressed by xing-qi plus a number from one to six. Sunday doesn’t use seven, but ri/tian (day). Formally, they can be written as XQ0-6. xing-qi is also called zhou (week) or li/bai (go to church) that is borrowed from religious activities. However, when we use zhou, Sunday cannot be said as *zhou-tian. Week days are usually placed after day and before hour as a parenthesis, such as 2013-10-15 (Tuesday) 3:00pm.

2.6 Festivals and Events

Some days or day sequences are named as festivals. Festivals are usually based on Gregorian system, such as the national day (guo-qing). In China, there are some festivals that are based on lunar system, such as the autumn day (zhong-qiu), which is Oct. 15th - Oct. 17th. When there is only one temporal element in the starting and ending time, which means that their parent elements are the same, the first time marker can be omitted. For example, brahman-biu wu-ri dao shi-yue shi-qi-ri (Oct. 15-17). Sometimes, only the length information is expressed, such as liang-nian (two years), which is made up of a Chinese number plus a classifier.

2.7 Referential time

The demonstrative, such as zhe (this) and na (that), can be placed before some temporal elements to form a referential time (ref). For example, zhe-yi-nian (this one year), ben-shi-ji (this century). The general pattern of such construction is [zhe/na]+[number]+[classifier]. There are also some lexicalized referential time expressions, such as jin-nian (this year), ming-tian (tomorrow) etc.

2.8 Durations

Duration is an interval of two time spots, i.e. the starting time and the ending time, connected by dao/zh (to). cong (from) can also be placed in front. For example, (cong) shi-yue shi-wu-ri dao shi-yue shi-qi-ri (Oct. 15th - Oct. 17th). When there is only one temporal element in the starting and ending time, which means that their parent elements are the same, the first time marker can be omitted. For example, shi-yue shi-wu-ri dao shi-qi-ri (Oct. 15-17). Sometimes, only the length information is expressed, such as liang-nian (two years), which is made up of a Chinese number plus a classifier.

2.9 Period phases

When talking about a specific time period, we can refer to its different phases, e.g. its starting period (chu-qi), middle period(zhong-qi) and final period (mo-qi/hou-qi). Period is different from duration...
The recognition of time expressions includes two phases: identify the temporal elements and then concatenate the elements to get sequences based on the topological relations of them and the constraints described in Table 1. The recognition of temporal elements are implemented by regular expressions. The topological relation could be modeled as an acyclic graph.

### 3.1 Convert to TIMEX3 format

In Chinese, the numbers in each temporal element can be a sequence of either Chinese or Arabic digits. For example, er-ling-ling-san-nian (year 2003) can also be written as 2003-nian. For this kind of expressions, we need a parser to get the Chinese numbers first, which has been embedded in our system. Meanwhile, it can also parse them into machine readable integers.

In Chinese, we can also use Arabic numbers. In our system, we build a parser that could translate both Arabic and Chinese number into machine readable integers. However, due to the space limitation, we will not describe the parser here. Once we get numbers for each element. Some heuristic rules can be used to filter some false positive examples. For instance, er-shi-san-dian (23:00) is a legal time expression, while er-shi-wu-dian (25:00) is illegal. It appears in text because it can also mean (25 points). We add constraint on the value of month(1, 12), day(1, 31), hour(0, 24) etc.

Based on our rule system, the converting to TIMEX3 format is quite straightforward since the rules are based on the inner structures of Chinese time entities. In cases of referential temporal elements, such as refyear, refday, we can first place a variable for further processing, since the resolution of such references is an independent task. However, this will be our future work. For festivals, as we mentioned that most festivals have fixed date. So, a festival dictionary will be needed.

Nevertheless, translating time entities into machine readable format is a great advantage of rule systems. Even though statical methods can give higher performance on recognition, there is no obvious way how to convert the time entities into machine readable format unless conversion rules are complied, which then will resort to the inner structure of the entities which is then the work done by our rule system.
Table 2: Corpus Information.

| Corpus       | #Words | #Entity |
|--------------|--------|---------|
| Sinica       | 10M    | 88K     |
| TempEval-2 Training | 23K    | 766     |
| TempEval-2 Test     | 10K    | 191     |

Table 3: Performance of the rule system on time entity extraction.

| Pattern                  | Prec. | #Rec. |
|--------------------------|-------|-------|
| month-xun                | 1.0   | 574   |
| month-day-dp-hour        | 1.0   | 356   |
| month-day-dp             | 1.0   | 315   |
| regnalyear-month-day     | 0.9985| 671   |
| month-day                | 0.9963| 7094  |
| refday-dp                | 0.9957| 2327  |
| year-month-day           | 0.9931| 2008  |
| regnalyear-month         | 0.9918| 363   |
| refyear-month            | 0.9910| 3098  |
| day-dp-hour              | 0.9875| 631   |
| dp-hour                  | 0.9855| 1764  |
| regnalyear               | 0.9836| 1319  |
| refyear-month-day        | 0.9831| 1221  |
| day-dp                   | 0.9831| 814   |
| refday-dp-hour           | 0.9824| 893   |
| year-month               | 0.9819| 1407  |
| year-season              | 0.9775| 261   |
| refday-dp-hour-minute    | 0.9755| 558   |
| century-periodphase      | 0.9674| 208   |
| season                   | 0.9658| 2401  |
| refday                   | 0.9622| 11670 |
| refyear                  | 0.9594| 3706  |
| century                  | 0.9473| 1384  |
| month                    | 0.9368| 4119  |
| dp-hour-minute           | 0.9336| 633   |
| year                     | 0.9247| 7324  |
| decade                   | 0.9148| 569   |
| weekday                  | 0.9147| 1458  |
| day                      | 0.8201| 3592  |
| hour-minute              | 0.8172| 474   |
| hour                     | 0.6673| 1073  |
| refyear-periodphase      | 0.4740| 219   |

4 Experiments

We use two different corpora: Sinica (Chen et al., 1996) and TempEval-2 from SemiEval-2010 competition (Pustejovsky and Verhagen., 2009). Sinica Corpus contains 10M words and the total number of time entity is 88K as shown in Table 2. The time words are tagged as ‘Nd’. However, there is no entity information. So, when an entity is recognized by our system, we first separate it into elements and then calculate the performance. Durations are labeled as number + classifier in Sinica, which are not time words. So, we don’t recognize durations in Sinica. For regnal year system, we only include a list of emperors of the Qing dynasty. We don’t deal with festivals as most of them are already lexicalized and are beyond the scope of entities. In other words, they can be recognized with a dictionary in a general word segmentation task.

TempEval-2 corpus includes training and test parts, as shown in Table 2. We analyse the annotation scheme based on training data and then add some additional rules on durations, such as shi-nian (ten years), shi-tian (ten days), and some approximate expressions, e.g. shi-ji-nian (more than then years) and so on. Meanwhile, we add three new elements: past (guo-qu), present (xian-zai), future (jiang-lai). Each element includes a list of Chinese words.

4.1 Experimental results and Discussion

Table 3 shows the overall performance on Sinica and TempEval-2 corpora. Our rule system gives a high performance. Table 4 shows the precision and the number of recalled entities for some selected frequent patterns from 91 patterns identified from Sinica. Some long patterns give 1.0 precision. Some patterns are quite ambiguous, such as hour-minute. This is due to fact that dian means both the point in float numbers and time hour, and fen means both minute and score point in Chinese. For example, san dian wu fen means both 3:05 and 3.5 points. Regarding the different performances of different patterns, we can assign a confidence value to each pattern, such as the length of the extracted patterns plus F1-value on a training corpus. This will be helpful when incorporating the patterns into other systems.

Basically, the longer the matched pattern is, the more confident it is. However, as we can see that, some long patterns have a low precision. This is
mainly due to the annotation errors that have split certain temporal elements into number-classifier construction in Sinica Corpus. For example, er-shi-wu-ri (the 25th) is annotated as er-shi-wu (25) plus ri (day).

Most ambiguous patterns contain one element, such as year and day. They can be both a date and a duration when the number is expressed in Chinese or Arabic digits. For example, 13 nian (13 year: year 2013) could also be thirteen years. In Sinica, durations are labeled as number + classifier, which are not time words. In TempEval-2 corpus, both date and duration are entities. So, it will not be a problem for detection on this corpus. The ambiguity of such patterns introduced most of the false positive examples.

Table 5 and Table 6 show the identified patterns and the precision and the number of recalled entities. Compared to Sinica corpus, TempEval-2 corpus is quite sparse, and the element refyear such as jin-nian (this year) and present such as mu-qian (currently), take up a large part of the entities. This problem will affect the evaluation result in that the identification of time words e.g. refyear and present will be important to the overall performance.

In order to compare our rule system with the state-of-the-art statistical models. We also built a

| Pattern        | Prec. | #Rec. |
|----------------|-------|-------|
| year           | 1.0   | 133   |
| month          | 1.0   | 8     |
| year-month     | 1.0   | 4     |
| decade-periodphase | 1.0 | 6     |
| refcentury-periodphase | 1.0 | 5     |
| refyear-firstmonth | 1.0 | 4     |
| refday-dp      | 1.0   | 4     |
| year-periodphase | 1.0 | 5     |
| refyear        | 0.9817| 107   |
| month-day      | 0.9783| 45    |
| refday         | 0.9412| 32    |
| refyear-periodphase | 0.9  | 9     |
| yearlength     | 0.875 | 56    |
| day            | 0.8571| 6     |
| year-month-day | 0.8571| 6     |
| present        | 0.848 | 106   |
| past           | 0.625 | 15    |

Table 5: Performance of the rule system on TempEval-2 training corpus.

| Pattern        | Prec. | #Rec. |
|----------------|-------|-------|
| refyear-month  | 1.0   | 9     |
| month-xun      | 1.0   | 2     |
| month-periodphase | 1.0 | 3     |
| year           | 1.0   | 14    |
| refyear-jd     | 1.0   | 3     |
| month-day      | 1.0   | 8     |
| refyear-periodphase | 1.0 | 18    |
| refyear        | 1.0   | 20    |
| present        | 0.9348| 43    |
| refyear-month  | 0.8571| 6     |
| future         | 0.8333| 5     |
| yearlength     | 0.625 | 10    |
| refday         | 0.6   | 3     |
| past           | 0.5714| 4     |

Table 6: Performance of the rule system on TempEval-2 test corpus.

Table 7: Features used in CRFs model.

| Type     | Feature                        |
|----------|--------------------------------|
| Context  | $token_{-1}$, $token_0$, $token_{1}$, $token_{-1} + token_0$, $token_0 + token_{1}$ |
| NGram    | unigram_of_token, bigram_of_token, trigram_of_token |
| Structure| end_with_classifier, start_with_number, number + classifier |

CRFs classifier with CRF++\(^1\) on TempEval-2 corpus. The features used are shown in Table 7. To study whether the rule system could help the statistical model, we also use the recognition results of our rule system as pattern features. The result is shown in Table 8. We can see that the rule system gives a much higher performance than CRFs without using the patterns as features, i.e. 0.8564 v.s. 0.7787.

\(^1\)http://crfpp.googlecode.com/svn/trunk/doc/index.html

we also conduct experiment to test the statistical model based on characters with features shown in Table 9. This setting is actually more reasonable than word based, since word segmentation and entity recognition are overlap tasks. The result is shown in Table 10. We can see that, compared to word based setting, the performance increased

\(^1\)http://crfpp.googlecode.com/svn/trunk/doc/index.html
### Table 8: Performance of time entity extraction with CRFs on TempEval-2 corpus.

| Feature | Precision | Recall | F1    |
|---------|-----------|--------|-------|
| Context | 0.7699    | 0.4555 | 0.5724|
| +Structure | 0.7867 | 0.6178 | 0.6921|
| +Ngram  | 0.8373    | 0.7277 | 0.7787|
| +Pattern| 0.8941    | 0.7958 | 0.8421|
| Rule System | 0.8876 | 0.8272 | 0.8564|

### Table 9: Features used in CRFs model based on characters.

| Type   | Feature                        |
|--------|--------------------------------|
| Context| `char-1`, `char0`, `char1`, `char-1+char0`, `char1+char1` |
| Structure| `is_number`, `is_classifier` |

### Table 10: Performance of time entity extraction with CRFs based on characters on TempEval-2 corpus.

|       | Precision | Recall | F1    |
|-------|-----------|--------|-------|
| CRFs  | 0.8476    | 0.7277 | 0.7831|
| +Pattern | 0.8977 | 0.8272 | 0.8610|
| Rule System | 0.8876 | 0.8272 | 0.8564|

from 0.7787 to 0.7831. This may due to the fact that with the segmentation information, the context features will be more sparse. When combining the patterns in CRFs model, the performance could be slightly improved. Overall, we can say that the inner structure of Chinese time entity is more important than context features.

The false negative examples of the rule system in Sinica includes some patterns that are not included in our system, some of which we think is not normal constructions of time expressions. For example, an Arabic digit sequence without the year marker `nian`, such as 2013, is also possibly a year element. Another one is the regnal year pattern, i.e. the `min-guo` period established in 1912 after Qing dynasty. However, there are many examples like `ba-shi-ba-nian` (88th years) with `min-guo` omitted.

The false negative examples of the rule system in TempEval-2 includes some time word-

5 Conclusion

In this paper, we made a linguistic study on Chinese time entities and presented a rule system for automatic recognition. We compare our system with CRFs model and the experiments on two different corpora showed that it gave a higher performance than the baseline system based on a CRFs model. When combining the rules with CRFs, the performance could be improved.

Acknowledgments

The work is supported by a General Research Fund (GRF) sponsored by the Research Grants Council (Project no. 543810 and 544011).

References

Yuen Ren Chao. 1968. *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

Keh-Jiann Chen, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. 1996. Sinica corpus: Design methodology for balanced corpora. In *Proceedings of Pacific and Asian Conference on Language and Information Computation*, pages 167–176.

James Pustejovsky and Marc Verhagen. 2009. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 112–116.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. Iso-timeml: An international standard for semantic annotation. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 18–21.

James Pustejovsky. 1995. *The Generative Lexicon*. Cambridge: The MIT Press.

Mingli Wu, Wenjie Li, Qin Lu, and Baoli Li. 2005. Ctemp: A chinese temporal parser for extracting and normalizing temporal information. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP05)*, pages 694–706.