Chinese Temporal Tagging with HeidelTime

Hui Li and **Jannik Strötgen** and **Julian Zell** and **Michael Gertz**
Institute of Computer Science, Heidelberg University
Im Neuenheimer Feld 348, 69120 Heidelberg, Germany
{hui.li,stroetgen,zell,gertz}@informatik.uni-heidelberg.de

Abstract

Temporal information is important for many NLP tasks, and there has been extensive research on temporal tagging with a particular focus on English texts. Recently, other languages have also been addressed, e.g., HeidelTime was extended to process eight languages. Chinese temporal tagging has achieved less attention, and no Chinese temporal tagger is publicly available. In this paper, we address the full task of Chinese temporal tagging (extraction and normalization) by developing Chinese HeidelTime resources. Our evaluation on a publicly available corpus – which we also partially re-annotated due to its rather low quality – demonstrates the effectiveness of our approach, and we outperform a recent approach to normalize temporal expressions. The Chinese HeidelTime resource as well as the corrected corpus are made publicly available.

1 Introduction

Temporal information plays a crucial role in many documents, and temporal tagging, i.e., the extraction of temporal expressions and their normalization to some standard format, is crucial for several NLP tasks. So far, research on temporal information extraction mostly focused on western languages, especially English. In contrast, eastern languages, e.g., Chinese, are less explored. Nevertheless, there is research on Chinese temporal tagging. While some works addressed either the extraction or the normalization subtask, a few full temporal taggers exist, e.g., CTEMP (Wu et al., 2005b) and CTAT (Jing et al., 2008), but none of them is publicly available.

In contrast, some temporal taggers were recently made available, e.g., DANTE (Mazur and Dale, 2009), TipSem (Llorens et al., 2010), and HeidelTime (Strötgen and Gertz, 2013). Furthermore, Strötgen et al. (2013) showed that HeidelTime can be extended to further languages by developing language-specific resources without modifying the source code. Thus, when developing temporal tagging capabilities for an additional language, one is faced with the question of whether to develop a new temporal tagger or to extend an existing one. We decided to extend HeidelTime for Chinese for the following reasons: (i) HeidelTime was the best temporal tagger in the TempEval-2 (English) and TempEval-3 (English and Spanish) competitions (Verhagen et al., 2010; UzZaman et al., 2013), (ii) it already supports eight languages, and (iii) it is the only multilingual temporal tagger for cross-domain temporal tagging, e.g., news- and narrative-style documents can be processed with high quality.

2 Related Work

For Chinese temporal tagging, machine learning and rule-based approaches have been employed. Wu et al. (2005a) and Wu (2010) report that machine learning techniques do not achieve as good results as rule-based approaches when processing Chinese. Thus, it is reasonable to extend a rule-based system such as HeidelTime to Chinese.

In general, temporal tagging approaches perform the extraction, the normalization, or both, and create TIDES TIMEX2 (Ferro et al., 2005) or TimeML's TIMEX3 (Pustejovsky et al., 2005) annotations. For development and evaluation, there are two Chinese temporally annotated corpora, the ACE 2005 training corpus and TempEval-2 (c.f. Section 3). Table 1 lists approaches to Chinese temporal tagging with some further information. The most recent work is the learning-based language-independent discriminative parsing approach for normalizing temporal expressions by Angeli and Uszkoreit (2013).

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 133–137,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

approach	tasks	method	standard	evaluation details	system available
Angeli and Uszkoreit (2013)	N	ML	TIMEX3	TempEval-2 (N)	no
Wu (2010)#	E	rules	TIMEX2	ACE 2007 (E)	no
Wen (2010)#	N	rules	TIMEX2	own corpus (N)	no
He (2009)#	E	ML+rules	TIMEX2	ACE 2005 (E)	no
Pan (2008)#	E	ML+rules	TIMEX2	ACE 2005 (E)	no
Jing et al. (2008)# – CTAT	E+N	ML+rules	TIMEX2	own corpus (E+N)	no
Wu et al. (2005b) – CTEMP	E+N	rules	TIMEX2	TERN 2004 (E), own corpus (E+N)	no
Hacioglu et al. (2005) – ATEL	E	ML+rules	TIMEX2	TERN 2004 (E)	no

Table 1: Information on related work addressing Chinese temporal tagging (#available in Chinese only).

There are also (semi-)automatic approaches to port a temporal tagger from one language to another. For instance, TERSEO (Negri et al., 2006; Saquete et al., 2006) has been extended from Spanish to English and Italian by automatic rule-translation and automatically developed parallel corpora. However, the normalization quality of this approach was rather low compared to a rule-based tagger manually developed for the specific language (Negri, 2007). This finding encouraged us to manually create Chinese HeidelTime resources instead of trying automatic methods.

3 The TempEval-2 Chinese Corpus

There are two Chinese temporally annotated corpora available: While the Chinese part of the ACE 2005 multilingual training corpus (Walker et al., 2006) has been used by some approaches (c.f. Table 1), it only contains TIMEX2 extent annotations. In contrast, the TempEval-2 Chinese data sets (Verhagen et al., 2010) contain TIMEX3 annotations with extent and normalization information. However, no TempEval-2 participants addressed Chinese and only Angeli and Uszkoreit (2013) report evaluation results on this corpus. Since HeidelTime is TIMEX3-compliant, and we address the extraction and normalization subtasks, we use the TempEval-2 corpus in our work.

3.1 Annotation Standard TimeML

For temporal expressions, TimeML (Pustejovsky et al., 2005) contains TIMEX3 tags with several attributes. The two most important ones – also annotated in the TempEval-2 data – are *type* and *value*. Type specifies if an expression is a date, time, duration, or set (set of times), and value contains the normalized meaning in standard format.

3.2 Original TempEval-2 Corpus

The Chinese training and test sets consist of 44 and 15 documents with 746 and 190 temporal expressions, respectively. However, several expressions have no normalized value information (85 in the training and 47 in the test set), others no type.[1]

This issue was also reported by Angeli and Uszkoreit (2013). Thus, they report evaluation results on two versions of the data sets, the original version and a cleaned version, in which all expressions without value information were removed.

3.3 Re-annotation of the TempEval-2 Corpus

Due to the significant amount of temporal expressions with undefined value attributes, we decided to manually assign normalized values to these expressions instead of excluding them. During this process, we recognized that the corpus contained several more errors, e.g., some expressions were annotated as dates although they refer to durations. Thus, instead of only substituting undefined values, we checked all annotations in the two data sets and corrected errors. For this, one Chinese native and two further TimeML experts discussed all modified annotations. Although there were several difficult expressions and not all normalizations were straightforward, we significantly improved the annotation quality. After our modification, the improved training and test sets contain 765 and 193 temporal expressions with value information, respectively. In Table 2, statistics about the three versions of the data sets are provided.

4 Chinese HeidelTime Resources

HeidelTime is a cross-domain, multilingual temporal tagger that strictly separates the source code and language-dependent resources (Strötgen and Gertz, 2013). While the implementation takes care of domain-dependent normalization issues, language-dependent resources contain pattern, normalization, and rule files. We had to develop such Chinese resources to perform Chinese temporal tagging with HeidelTime.

[1] Note that the TempEval-2 corpus developers stated that the annotations of the non-English documents are rather experimental (Verhagen, 2011).

corpus	docs	temp. expr.	date / time / duration / set	undef value
training set				
original	44	746	623 / 10 / 113 / 0	85
AU13-clean	44	661	555 / 10 / 96 / 0	0
improved	44	765	628 / 10 / 125 / 2	0
test set				
original	15	190	160 / 0 / 27 / 3	47
AU13-clean	15	143	128 / 0 / 15 / 0	0
improved	15	193	166 / 0 / 23 / 4	0

Table 2: Statistics on the three versions of the Chinese TempEval-2 data sets.

4.1 Chinese Linguistic Preprocessing

As input, HeidelTime requires sentence, token, and part-of-speech information. For most of the supported languages, HeidelTime uses a UIMA wrapper of the TreeTagger (Schmid, 1994). Since there is also a Chinese model for the TreeTagger available, we rely on the TreeTagger for Chinese linguistic preprocessing.[2]

4.2 Resource Development Process

To develop Chinese HeidelTime resources, we followed the strategy applied by Strötgen et al. (2013) for Spanish: Using HeidelTime's English resources as starting point, we translated the pattern files, the normalization files, and the rules for extracting and normalizing temporal expressions. More details on these steps are provided next.

Pattern & Normalization Resources. English patterns in the pattern files, which also exist in Chinese in a similar form, were directly translated. For instance, there are Chinese expressions for names of months and weekdays. Patterns existing in English but not used in Chinese were removed, e.g., there are no abbreviations of month names in Chinese. In contrast, for other patterns frequently used in Chinese, additional pattern files were created. Examples are Chinese numerals.

Based on the pattern files, we built the normalization files. Here, the normalized values of the patterns are stored. An example of the Chinese resources is as follows: The three patterns "星期二", "礼拜二", and "周二" can all be translated as Tuesday and are thus part of the Weekday pattern resource. Since weekdays are internally handled by HeidelTime with their English names, the normalization file for Chinese weekdays contains "星期二,Tuesday" "礼拜二,Tuesday" and "周二,Tuesday".

Chinese Rule Development. HeidelTime's rules contain three components, a name, an extraction and a normalization part. The extraction mainly makes use of regular expressions and the pattern resources, and in the normalization part, the matched expressions are normalized using the normalization resources.[3] To develop the rules, we again followed Strötgen et al. (2013) and applied the following strategy:

(i) A few simple Chinese rules were created based on the English rules. (ii) We reviewed extracted temporal expressions in the training set and improved the extraction and normalization parts of the rules. (iii) We checked the training texts for undetected expressions and created rules to match them. In parallel, we adapted the Chinese pattern and normalization resources when necessary. (iv) We translated more complex English rules to also cover valid expressions not occurring in the Chinese training documents. (v) Steps (ii) to (iv) were iteratively performed until the results on the training set could not be improved further.

4.3 Chinese Challenges

Chinese is an isolating language without inflection and depends on word order and function words to represent grammatical relations. Although we only consider modern Mandarin as it is the most widely used variety of Chinese in contemporary texts, many challenges occurred during the resource development process. Some examples are:

Polysemous words: Many Chinese words have more than one meaning, e.g., dynasty names such as "唐" (Tang) or "宋" (Song) can refer to a certain time period, but also appear as family names.

Further ambiguities: There are many ambiguous expressions in Chinese, e.g., the temporal expression "五日前" has two meanings: "before the 5th day of a certain month" and also "5 days ago" – depending on the context.

Calendars: There are various calendars in Chinese culture and thus also in Chinese texts, such as the lunar calendar and the 24 solar terms, which are different from the Gregorian calendar and thus very difficult to normalize. Besides, Taiwan has a different calendar, which numbers the year from the founding year of the Republic of China (1911).

[2] `http://corpus.leeds.ac.uk/tools/zh/`.

[3] For more details about HeidelTime's system architecture and rule syntax, we refer to Strötgen and Gertz (2013).

training set	P	R	F	value	type
original	96.1	92.7	94.4	80	93
AU13-clean	80.7	95.1	87.3	91	95
improved	97.6	94.4	96.0	92	95
test set	P	R	F	value	type
original	93.4	82.0	87.3	70	93
AU13-clean	63.5	88.0	73.8	89	96
improved	95.5	83.8	89.3	87	96

Table 3: Evaluation results for extraction and normalization (TempEval-2 training and test sets).

5 Evaluation

In this section, we present evaluation results of our newly developed Chinese HeidelTime resources. In addition, we compare our results for the normalization sub-task to Angeli and Uszkoreit (2013).

5.1 Evaluation Setup

Corpus: We use three versions of the TempEval-2 training and test sets: (i) the original versions, (ii) the improved versions described in Section 3.3, and (iii) the cleaned versions also used by Angeli and Uszkoreit (2013) in which temporal expressions without value information are removed.

Setting: Since the TempEval-2 data already contains sentence and token information, we only had to perform part-of-speech tagging as linguistic preprocessing step. For this, we used the TreeTagger (Schmid, 1994) with its Chinese model.

Measures: We use the official TempEval-2 evaluation script. For the extraction, precision, recall, and f-score are calculated on the token-level. For the normalization, accuracy for the attributes *type* and *value* are calculated on the expression-level. Note that the use of accuracy makes it difficult to compare systems having a different recall in the extraction, as will be explained below.

5.2 Evaluation Results

Table 3 (top) shows the evaluation results on the training set. Extraction and normalization quality are high, and value accuracies of over 90% on the cleaned and improved versions are promising.[4]

The results on the test sets (Table 3, bottom) are lower than on the training sets. However, value accuracies of almost 90% with a recall of more than 80% are valuable and comparable to state-of-the-art systems in other languages. A first error analysis revealed that while the training documents are written in modern Mandarin, some test documents contain Taiwan-specific expressions (c.f. Section 4.3) not covered by our rules yet.

	original		AU13-clean		# correct
training set	value	type	value	type	value
AU13	65%	95%	73%	97%	484[5]
HeidelTime	80%	93%	91%	95%	574
	original		AU13-clean		# correct
test set	value	type	value	type	value
AU13	48%	87%	60%	97%	86[5]
HeidelTime	70%	93%	89%	96%	121

Table 4: Normalization only – comparison to AU13 (Angeli and Uszkoreit, 2013).

Finally, we compare the normalization quality of our approach to the multilingual parsing approach of Angeli and Uszkoreit (2013). However, their approach performs only the normalization subtask assuming that the extents of temporal expressions are provided. For this, they used gold extents for evaluation. HeidelTime only normalizes those expressions that it knows how to extract. Thus, we run HeidelTime performing the extraction and the normalization. However, since the accuracy measure used by the TempEval-2 script calculates the ratio of correctly normalized expressions to all extracted expressions and not to all expressions in the gold standard, we additionally present the raw numbers of correctly normalized expressions for the two systems. Table 4 shows the comparison between our approach and the one by Angeli and Uszkoreit (2013). We outperform their approach not only with respect to the accuracy but also with respect to the numbers of correctly normalized expressions (574 vs. 484[5] and 121 vs. 86[5] on the training and test sets, respectively) – despite the fact that we perform the full task of temporal tagging and not only the normalization.

6 Conclusions & Ongoing Work

In this paper, we addressed Chinese temporal tagging by developing Chinese HeidelTime resources. These make HeidelTime the first publicly available Chinese temporal tagger. Our evaluation showed the high quality of the new HeidelTime resources, and we outperform a recent normalization approach. Furthermore, the re-annotated Chinese TempEval-2 data sets will also be made available.

Currently, we are performing a detailed error analysis and hope to gain insights to further improve HeidelTime's Chinese resources.

[4] Note that the lower value accuracy on the original set is due to expressions without value information in the gold standard, and that the low extraction precision in the clean version is due to some of those expressions being (correctly) extracted by the system but removed from the gold standard.

[5] Number of correct value normalizations calculated based on value accuracy and number of expressions in the data sets.

References

Gabor Angeli and Jakob Uszkoreit. 2013. Language-Independent Discriminative Parsing of Temporal Expressions. In *P*roceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), pages 83–92.

Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth Sundheim, and George Wilson. 2005. TIDES 2005 Standard for the Annotation of Temporal Expressions. Technical report, The MITRE Corporation.

Kadri Hacioglu, Ying Chen, and Benjamin Douglas. 2005. Automatic Time Expression Labeling for English and Chinese Text. In *P*roceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2005), pages 548–559.

Ruifang He. 2009. *Research on Relevant Techniques of Temporal Multi-document Summarization*. Ph.D. thesis, Harbin Institute of Technology.

Lin Jing, Cao Defang, and Yuan Chunfa. 2008. Automatic TIMEX2 Tagging of Chinese Temporal Information. *J*ournal of Tsinghua University, 48(1):117–120.

Hector Llorens, Estela Saquete, and Borja Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. In *P*roceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010), pages 284–291.

Pawel Mazur and Robert Dale. 2009. The DANTE Temporal Expression Tagger. In *P*roceedings of the 3rd Language and Technology Conference (LTC 2009), pages 245–257.

Matteo Negri, Estela Saquete, Patricio Martínez-Barco, and Rafael Muñoz. 2006. Evaluating Knowledge-based Approaches to the Multilingual Extension of a Temporal Expression Normalizer. In *P*roceedings of the Workshop on Annotating and Reasoning about Time and Events (ARTE 2006), pages 30–37.

Matteo Negri. 2007. Dealing with Italian Temporal Expressions: The ITA-CHRONOS System. In *P*roceedings of EVALITA 2007, pages 58–59.

Yuequn Pan. 2008. Research on Temporal Information Recognition and Normalization. Master's thesis, Harbin Institute of Technology.

James Pustejovsky, Robert Knippen, Jessica Littman, and Roser Sauri. 2005. Temporal and Event Information in Natural Language Text. *L*anguage Resources and Evaluation, 39(2-3):123–164.

E. Saquete, P. Martínez-Barco, R. Muñoz, M. Negri, M. Speranza, and R. Sprugnoli. 2006. Multilingual extension of a temporal expression normalizer using annotated corpora. In *P*roceedings of the EACL 2006 Workshop on Cross-Language Knowledge Induction, pages 1–8.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *P*roceedings of the International Conference on New Methods in Language Processing.

Jannik Strötgen and Michael Gertz. 2013. Multilingual and Cross-domain Temporal Tagging. *L*anguage Resources and Evaluation, 47(2):269–298.

Jannik Strötgen, Julian Zell, and Michael Gertz. 2013. HeidelTime: Tuning English and Developing Spanish Resources for TempEval-3. In *P*roceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), pages 15–19.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *P*roceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), pages 1–9.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *P*roceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010), pages 57–62.

Marc Verhagen. 2011. TempEval2 Data – Release Notes. Technical report, Brandeis University.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 Multilingual Training Corpus. Linguistic Data Consortium, Philadelphia.

Yanxia Wen. 2010. Research on Time Standardization in Chinese. Master's thesis, Shanxi University.

Mingli Wu, Wenjie Li, Qing Chen, and Qin Lu. 2005a. Normalizing Chinese Temporal Expressions with Multi-label Classification. In *P*roceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE 2005), pages 318–323.

Mingli Wu, Wenjie Li, Qin Lu, and Baoli Li. 2005b. CTEMP: A Chinese Temporal Parser for Extracting and Normalizing Temporal Information. In *P*roceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP 2005), pages 694–706.

Tong Wu. 2010. Research on Chinese Time Expression Recognition. Master's thesis, Fudan University.