社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－NL－179 (11)
2007－SLP－ 66 (11)
2007／5／25

依存構造を用いた中国語事象の時間関係のタグ付きコーパスの構築

鄭育昌　浅原正幸　松本裕治
奈良先端科学技術大学院大学　情報科学研究科
{yuchan-c, masayu-a, matsu}@is.naist.jp

概要：本稿では中国語新聞記事に事象間の時間関係のタグ付け作業を説明するガイドラインを示す。本研究の目的は機械学習で事象間の時間関係を自動的に解析するため、事象間の時間関係のタグ付きコーパスを構築することである。文章の全ての事象の組合せを考慮する解析は非効率であるため、依存構造の導入によって、内容の理解に重要な事象間時間関係を同定する手法を提案する。本手法の有効性を図るため、本稿では小規模コーパスを人手で構築し、本手法の有効性を検証した。依存構造の導入により、事象間の可能な時間関係が約 49%再現でき、コーパス構築するための作業量が減少できることが判明した。

Constructing a Temporal Relation Tagged Corpus of Chinese based on Dependency Structure Analysis

Yuchang CHENG, Masayuki ASAHARA, Yuji MATSUMOTO
Graduate School of Information Science
Nara Institute of Science and Technology
{yuchan-c, masayu-a, matsu}@is.naist.jp

Abstract: This paper describes an annotation guideline for a temporal relation tagged corpus. Our goal is to construct a machine learnable model that automatically analyzes temporal events and relations between events. Since analyzing all combinations of events is inefficient, we examine use of dependency structure analysis to efficiently recognize meaningful temporal relations. We survey a small tagged data set to investigate the coverage of our method. Although the coverage of our methods is about 49%, we find that the dependency structure appears useful for reducing manual efforts in constructing a tagged corpus with temporal relations.

1. Introduction

Extracting the temporal information in articles is useful technique for many NLP applications such as question-answering, text summarization, machine translations and so on. The temporal information includes three parts: 1. temporal expressions, which describe time or period in the real world; 2. events, which are situations that occur or happen, punctual or lasting for a period of time; 3. temporal relative relations, which describe the relative relation between an event and a temporal expression, or between two events.

There are many researches dealt with the temporal expressions and events. Extracting temporal expressions is a subtask of NER and widely studied in many languages. Normalizing the temporal expressions is also investigated in evaluation workshop. Event semantics are also investigated in linguistics and AI fields. However, researches on temporal relation extraction are still limited. The temporal relation extraction includes the following issues: identifying events, anchoring an event in time, ordering events and reasoning with contextually underspecified temporal expressions. TimeBank [Pustejovsky 06] can be used for developing machine learning approaches to automatically extract and recognize the temporal relation in English. There is no publicly available resource for the temporal information processing in Chinese. We aim to efficiently construct a temporal relation tagged corpus of Chinese for developing a temporal relation analyzer.

This paper presents how efficiently construct temporal relation tagged corpus of Chinese. First, we describe a guideline of corpus annotation. Our annotation guideline is based on TimeML [Saurí 05] which is originally for English texts. Second, we propose use of dependency structure, which reduces manual efforts. The dependency structure helps to detect subordinate structures of the sentence. Third, we investigate distribution of the temporal relation in Chinese news texts.

Temporal relation includes the anchoring relation from an event to a temporal expression, and the ordering relation between two events. We focus on the ordering relations in this article.

2. TimeML: an annotation guideline

TimeBank is a temporal information tagged corpus that includes full temporal information. The corpus is annotated by the TimeML standard. Table 1 lists the definition of the tags. "EVENT", "TIMEX3" and "SIGNAL" tags in TimeML mark up the temporal entities such as events and temporal expressions. Link tags annotate the temporal relations between entities. The definitions of temporal relations with the tag "TLINK" are based on Allen's [Allen 83] temporal relations. The tag "SLINK" and "ALINK" annotate the relations between a main event and its subordinate event. Whereas the tag "ALINK" describes an aspectual relation, the tag "SLINK" describes a subordinate relation without explicit aspectual meaning.

We refer to the TimeML languages to define our standard of Chinese temporal relation tagged corpus. TimeBank include all possible relations between an event and a temporal expression or between two events, but we only consider the relations between two events. TimeBank is tagged manually and extracted all information that can be understood in the English articles. We wanted to construct a Chinese temporal relation tagged corpus similar to English TimeBank but it will base on dependency structure for reducing efforts.

3. The temporal relation annotation based on dependency structure

We only annotate the temporal relation between events by verbs. When an article includes n events, we need to annotate $_nC_2$ event pairs. It is less certain that a long distance event pair has a temporal relation because most long distance event pairs have no direct relation. Annotating all event pairs is inefficient; therefore we want to use less human effort to extract more meaningful relations. Thus, we annotate the following event pairs: 1. adjacent event pairs in the document, 2. the head-modifier event pairs in a dependency structure, 3. the sibling event pairs in a dependency structure. After extracting the temporal relations from the dependency structure, we adopt transitive rules to extend the relations.

Tags	EVENT	MAKEINSTANCE	TIMEX3	SIGNAL
Number	7935	7940	1414	688
	All links	adjacent links	head-modifier links	
Tlink	6418	1757	1186	
Slink	2932	2129	2174	
Alink	265	167	157	
Total	9615	4053	3517	

Table 1: Distribution of tags in TimeBank

3.1 Data analysis of TimeBank

TimeBank 1.2[1] contains 183 articles with just over 61,000 non-punctuation tokens. We investigate the distribution of temporal tags as shown in Table 1. TimeBank includes 9615 links (TLINK, SLINK, and ALINK). Of them, 4053 links are the relation between adjacent entity pairs[2]. According to the distribution, if we are able to recognize more adjacent relations correctly, we expect that the adjacent relations are extended by transitive rules cover more than 50% of the total relations in the corpus. To recognize the adjacent links of events, we only annotate all event pairs that are adjacent (the adjacent pair means the focus event and its linearly preceding event).

Additionally, we can find that about 50% of the adjacent links is SLINK. The tag "SLINK" means a subordinate relation between events but not a temporal relation. This observation gives us the idea that to recognize SLINKs is an important task for extracting adjacent relations.

3.2 Adjacent event pairs in Chinese article

An example phrase "停止拨付财政债券安排的资金并起诉 (*To stop providing funds that were prepared by financial bond, and to prosecute...*) in Fig. 1 has four events: "停止 (*stop*)", "拨付 (*provide*)", "安排 (*prepare*)" and "起诉 (*prosecute*)". The temporal order of these events is shown in the lower part of Fig. 1. We can get six meaningful temporal relations from this example, and the relations are listed in Fig. 1.

The linear adjacent pairs of these events are {停止-拨付, 拨付-安排, 安排-起诉}, and we can extract the temporal relation of these events and extend the relations using transitive rules. However, the relation of adjacent event pair "安排-起诉" is not useful information for readers because the event "安排 (*prepare*)" is a subordinate event of the event "拨付(*provide*)." The temporal relation between events "停止 (*stop*)" and "起诉 (*prosecute*)" is more useful than the relation between events "安排 (*prepare*)" and "起诉 (*prosecute*)" because events "停止 (*stop*)" and "起诉 (*prosecute*)" are coordinate events. It

1 http://www.ldc.upenn.edu/

2 The tag "TLINK" includes the temporal relations between document time and other temporal entities in an article, and includes the temporal relations between two matrix verb events of different sentences.

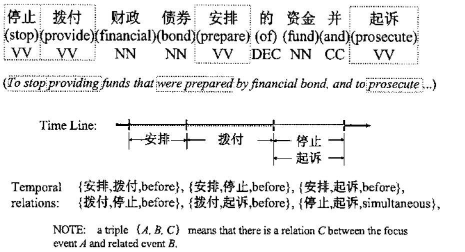

Fig. 1: an example sentence with the temporal order of events

should be noted that the subordinate relations do not include temporal relation in TimeML. However, our empirical observation finds that many subordinate event pairs can include temporal relations, such as the two events "安排 (*prepare*)" and "拨付 (*provide*)". Our criteria require annotators to recognize the temporal relation of subordinate event pairs as much as they can.

In this example, a native annotator can understand the temporal relation between "安排 (*prepare*)" and "起诉 (*prosecute*)" is "before[3]". However, many event pairs are similar to this example do not have an explicit temporal relation. Either in this case, the cognitive process by which the native reader understands the relation of the event pair "安排 (*prepare*)" and "起诉 (*prosecute*)" is as follows. First, the event "安排 (*prepare*)" occurs before the event "拨付 (*provide*)", and the event "拨付 (*provide*)" occurs before the event "停止 (*stop*)". Second, the two events "停止 (*stop*)" and "起诉 (*prosecute*)" are coordinated and occur at the same time. Therefore, the event "安排 (*prepare*)" should occur before the event "起诉 (*prosecute*)". To analyze this kind of event pairs ("安排 (*prepare*)" and "起诉 (*prosecute*)"), we should consider not only the adjacent observation of events but also the syntactic structure of a sentence to acquire the correct temporal information.

3.3 The head-modifier and sibling event pair on dependency structure

The reason that we adopt the dependency structure to extract the temporal relation is that the dependency structure can describe the head-modifier relation between words. We define the verbs as the events and we only focus on the relation between verbs in a dependency structure. The dependency structure of the example sentence in Fig. 1 is illustrated in Fig. 2, the upper arrows on words point to their head word.

[3] All temporal relations are in Fig. 4.

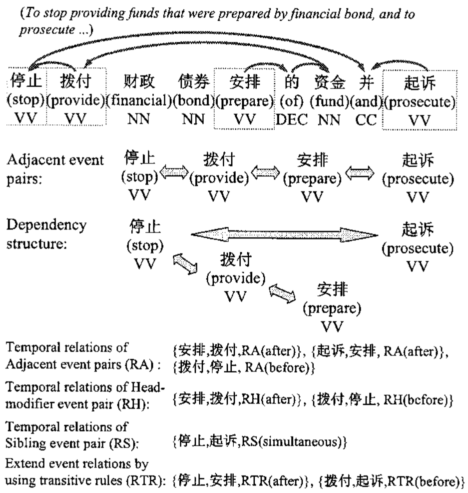

Fig. 2: an example sentence with dependency structure

According to our empirical observation in the Penn Chinese Treebank, most sentences include more than one verb. Many of the verbs modify another verb in a dependency structure and can be regarded as subordinate events. Therefore, to annotate the temporal relation of these head-modifier event pairs is just as important as of the adjacent pairs.

The punctuation "," usually be used in the semantic ending of a sentence in Chinese. To distinguish the meaning of the punctuation mark "," is difficult. The average length of sentences in the Penn Chinese Treebank is 27 words. Therefore a sentence in treebank could include several clauses which denote independent events. Although the definition of a Chinese "sentence" is ambiguous, we recognize that a sentence is ended by the punctuation "。" (a full stop). For extracting the temporal relations of the event pairs between different clauses in a similar sentence, it is necessary to analyze the relations of sibling event pairs.

In the example sentence, the event "安排 (*prepare*)" modified the event "拨付 (*provide*)", and the event "拨付 (*provide*)" modified the event "停止 (*stop*)". We can determine these head-modifier event pairs as subordinate relations. For the event "起诉 (*prosecute*)", the most important information is the relation between the event pair "停止 (*stop*)" and "起诉 (*prosecute*)" because this event pair is a coordinated event pair. We define the event pair that shares a head event as a sibling event pair. The coordinated event pair "停止 (stop)" and "起诉 (prosecute)" is defined as a sibling event pair.

	RA+RH+RS	RA	RH+RS	w/o transitive rules
Relations of Adjacent event pair (RA)	772	772	0	772
Relations of Head-modifier event pair (RH)	644	0	644	644
Relations of Sibling event pair (RS)	215	0	215	215
Total extracted event relations	1151	772	859	1151
Extend event relations by using transitive rules	3605	2071	2755	1151
All event combinations	5164	5164	5164	5164
CORRECT relations	2543	1404	2046	944
Precision	0.705	0.677	0.74	0.820
Recall	0.492	0.271	0.396	0.182

Table 2: Results of the preliminary survey

In our corpus, we annotate the temporal relation of all head-modifier event pairs and the sibling events according to the dependency structure of the sentence except the adjacent event pairs, and annotate the subordinate relation of the head-modifier event pairs (if it is subordinate pair). After annotating these relation tags, we use transitive rules to extend the temporal relations.

The below of Fig. 2 describes the tagging process of our method. After extracting the temporal relations of adjacent event pairs, head-modifier event pairs and sibling event pairs, using transitive rules can acquire new relations "{停止, 安排, RTR(after)}, {拨付, 起诉, RTR(before)}." We do not need analyze all possible event pairs and can acquire many useful temporal relations by our method.

4. Coverage of the limited links

Before we annotate the corpus, we should evaluate the coverage of our supposed criteria. We investigate a small corpus to observe the performance of our criteria. For n events in an article, there are ${}_nC_2$ relations that should be considered. We can compare the relations of all pairs of events and the relations extracted by our criteria to observe the coverage of our criteria. However, it is difficult to annotate the temporal relations of all event pairs. We select 30 articles and only use the first two paragraphs of each article to make our survey data. The small corpus includes 429 events and 3092 tokens.

We annotate the small corpus manually both by extracting all event pairs and by using our criteria. After annotating by our criteria, we use simple transitive rules to extend the relations. For example, if two temporal relations "Event A occurs during Event B" and "Event B occurs before Event C" are extracted, we can infer a new relation "Event A occurs during Event C". In preceding research [Mani 2006], the transitive rules could adopt some syntactic or semantic features of the event pair to extend more transitive rules. In this paper, we use simple transitive rules that only adopt the unambiguous relations and without syntactic / semantic feature.

We compare four methods to extract temporal relations. The methods are: 1. Using relations of adjacent event pairs, head-modifier event pairs and sibling event pairs, then to extending the relations by transitive rules (The column "RA+RH+RS" in Table 2). 2. Only using the relations of adjacent event pairs with transitive rules (The column "RA" in Table 2). 3. Using the relations of head-modifier event pairs and sibling event pairs with transitive rules (The column "RH+RS" in Table 4). 4. Using three kinds of event pairs without transitive rules (The column "w/o transitive rules" in Table 2). For experimental convenience, we reduce nine classes of temporal relations to five classes. The classes {after, overlap_by, begun_by} are reduced to the class "after" and the classes {before, overlap, ended_by} are reduced to the class "before."

Table 2 describes the distribution and coverage of our proposed methods. We regard the relations of all event pairs as the gold standard (the row "All event combinations") and we compare the result of our method with the gold standard. The row "Recall" shows the coverage of each method and the row "Precision" shows the accuracy of our method.

The last column shows the case of using our criteria to extract temporal relations without using transitive rules. The row "Extend event pair relations using transitive rules" in this column indicates the total amount of events that are extracted by our criteria. It should be noted that an adjacent event pair could be also a sibling event pair or a head-modifier event pair, therefore the number of the relations that we extract by our criteria is not equal to the total number of the three kinds of relation types (RA+RH+RS > Total event pairs).

Intuitively, the combination of events must include all relations that could be extracted. The relations that we extract by our criteria must be included in the gold standard. However, the "Extend event relations by using transitive rules" row of the "Without transitive rules" is not 100% included in all combination of events (The "Recall" is less than 1). This performance of recall can be thought as the limit of tagging consistency when the annotators are working independent. One reason of this observation is that the annotator does not consider any syntactic structure in annotating the event combination. The intuitive reorganization of event relations could be inconsistent with the dependency structure. This observation indicates the difficulty of constructing a corpus con-

Attribute	values	definition
the temporal properties of the event		
E-dynamic	State, dynamic	Activity of event
E-period	Durative, instantaneous, forever	Period of event
E-telicity	Telic, non-telic	Telicity of event
the temporal relation tag of the event		
Rel-linear-preceding	Relations in Fig. 4 + first, unknown, passive	Relation between the focus event and the linear adjacent preceding event
Rel-tree-preceding	Same as upper row	Relation between the focus event and the sibling event
Rel-tree-ancestor	Same as upper row	Relation between the focus event and the ancestor event
Sub-ord	modal, explanation, condition, none, report	Subordinate type between the focus event and the ancestor event
information of the main verb		
ancestor-verb		The ancestor verb of the main verb of the event
eventid		the ID of the event
maindep		the head word ID of the focus word
mainid		the ID of the main verb
mainpos		the POS tag of the main verb
mainword		the main verb

Table 3: The attributes of an event

sistently. To retain the consistency in our full corpus, we should repeat the annotation and check the data by a different annotator.

According our results, the precision of using "RH+RS" with transitive rules is better than the one of only using "RA" with rules. The head-modifier event pairs can extract many important relations that the adjacent event pair cannot extract. The recall row shows the coverage of our method. Although we use three types of event pairs and transitive rules, only the 49% relations of the gold standard can be extracted. We can add more transitive rules that consider other syntactic or semantic information of events to extend the relations.

Attribute	Values	Number
E-dynamic	State / dynamic	5347 / 1892
E-period	Durative/ instantaneous/ forever	3024/4156/59
E-telicity	Telic / non-telic	3440 / 3799
Rel-linear-preceding	(top four relations) After / simultaneous / before / during	2523 / 2065 / 1091 / 463
Rel-tree-preceding	(top four relations) None / after / simultaneous / before	5116 / 818 / 491 / 305
Rel-tree-ancestor	(top four relations) None / simultaneous / before / after	1968 / 1816 / 1773 1073
Sub-ord	Total subordinate relations	3422

Table 4: The results of the attributes

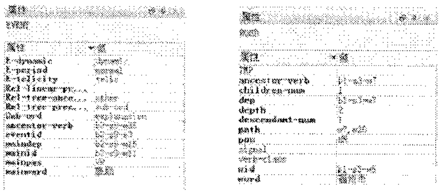

Fig. 3: Attribute windows for annotators.

5. Constructing the corpus

5.1 Basic data

To recognize the subordinate event pairs and parent-child event pairs, we needed a dependency parsed corpus. We used Penn Chinese TreeBank [Palmer 06] as our original data. However, Penn Chinese TreeBank do not include the modifier-head relations, we translated phrase structures to dependency structures by using head rules [Cheng 2005].

5.2 Data format and Temporal relation

The definition of event in TimeML includes verbs, predicative clauses, nominalizations...etc. But researchers usually narrow down the definition of event to verbs because world knowledge is necessary for extracting other types of events. We also define event as verb in our standard. We tagged the three types (adjacent event pairs, head-modifier event pairs and sibling event pairs) of event pairs manually. The annotator would decide most appropriate relation of these types of each event. Fig. 3 shows the attribute information that annotators view. The right side window describes the morphological information and the dependency information of a word, such as that the column "dep" means the head word ID of the focus word and the column "ancestor-verb" shows the upper verb in the dependency tree.

The left side window in Fig. 3 shows attributes of the focus event. Table 3 describes attributes of an event and that are what we required annotators to do. The attributes of an event roughly include two parts: Properties of event (E-dynamic, E-period, and E-telicity) and temporal relations (Rel-linear-preceding, Rel-tree-preceding and Rel-tree-ancestor). Annotators should decide the appropriate selection for each attribute. Properties of event are the temporal characteristic of event; these characteristic roughly correspond to the classification of verbs in [Dorr 1997]. We do not require annotators to classify events to several verb classes, but instead of three binary selections. The possible temporal relations are shown in Fig. 4, which compared our standard to TimeML and Allen's definition. Our definition of temporal relations is based on TimeML language and Allen's research. In Fig. 4, EVENT 1 is the focus event and EVENT 2 is the related event. The "Sub-ord" column in Table 3 means the subordinate relation between events and we refer to TimeML to define the subordinate relations.

5.3 Progress and future direction

(1) Progress

The Penn Chinese Treebank 5.0 contains 507,222 tokens, 18,782 sentences, and 890 articles. We will automatically analyze these attributes in the future, but we need a manually tagged training data to construct machine learning models. We use a part of the Penn Chinese TreeBank (about 10%) to construct a basic data set. Because the inconsistency of the larger corpus might exist in this annotated corpus, we could not train it to get machine learning models before we repeat the annotating work.

Some results of the training data by hand are summarized in Table 4. Because the temporal relations have more than ten types, we only show the top four relations and only show the total number of subordinate relations. Considering the tag "Rel-linear-preceding (adjacent event pairs)", the relation classes "After / simultaneous / before" are the most possible relation of adjacent event pairs. Because we request the annotators to annotate the temporal relations as possible, they used much world knowledge and the information in other parts of the article. Therefore the class "unknown" in tag "Rel-linear-preceding" is infrequent. The relation class "none" of the tag "Rel-tree-preceding (sibling event pairs)" means the focus event does not have any sibling event because events in similar sentences are structured as a hierarchy structure and there are few events that modify same head events. Therefore, most events are singletons of their head events. In the tag "Rel-tree-ancestor (head-modifier event pairs)," the root event of the dependency structure does not have a head event and the correct selection of the tag "Rel-tree-ancestor" in this case should be "none". In the tag "sub-ord (subordinate relation)," most types of subordinate relation are explanations; therefore, we only show the total number of subordinates in the data.

(2) Future direction

To construct such temporal relation tagged corpus is cumbersome. Although the events can be identified automatically, the working time of each article is 50 minutes. However, we can extend the extracted relations by using induction rules such as: if event A occurs before event B and event B occurs before event C, than event A occurs before event C…etc. After that, we will use this small data as training data for machine learning, then tagging the attributes of events automatically.

	Our criterion	TimeML	Allen
EVENT 1 EVENT 2	AFTER	AFTER	after
EVENT 1 EVENT 2		IAFTER	met-by
EVENT 1 EVENT 2	OVERLAPPED−BY		overlapped-by
EVENT 1 EVENT 2		ENDS	finishes
EVENT 1 EVENT 2	DURING	DURING/IS_INCLUDED	during
EVENT 1 EVENT 2	BEGUN_BY	BEGUN_BY	started-by
EVENT 1 EVENT 2	SIMULTANEOUS	SIMULTANEOUS/IDENTITY	equal
EVENT 1 EVENT 2	INCLUDES	INCLUDES/DURING_INV	contains
EVENT 1 EVENT 2	ENDED_BY	ENDED_BY	finished-by
EVENT 1 EVENT 2	OVERLAPS		overlaps
EVENT 1 EVENT 2		BEGINS	starts
EVENT 1 EVENT 2	BEFORE	IBEFORE	meets
EVENT 1 EVENT 2		BEFORE	before

Fig. 4: the relation definition among our criteria, TimeML and Allen's work

6. Conclusion

This research focuses on an annotation guideline of temporal relation tagged corpus of Chinese. The guideline is based on the TimeML language but we adopt dependency structure information to acquire more meaningful temporal relations with less manual effort. We define events as the verbs and define three types of link for event pairs. These types (adjacent event pairs, head-modifier event pairs and sibling event pairs) include most meaningful information and can resolve the problem of subordinate relation. We use a part of Penn Chinese TreeBank to construct a small training data. In future, we will investigate machine learning approaches to tag annotation automatically and acquire the coverage of our results and the results of TimeML-like manual tagged corpus.

Reference

[Allen 83] James F. Allen, "Maintaining Knowledge about Temporal Intervals", Communications of the ACM Vol.26 No.11, 1983.

[Cheng 05] Yuchang Cheng, "Chinese Deterministic Dependency Analyzer: Examining Effects of Chunking, Root node finder and Global Features", Master thesis, NAIST, 2005.

[Dorr 97] Bonnie J. Dorr and Mari Broman Olsen, "Deriving Verbal and Compositional Lexical Aspect for NLP Application", ACL 1997, 1997.

[Mani 06] Inderjeet Mani, et al., "Machine Learning of Temporal Relations", COLING/ACL 2006, 2006.

[Palmer 05] Martha Palmer, et al., "Chinese Treebank 5.1" http://www.ldc.upenn.edu/, LDC, 2005.

[Pustejovsky 06] James Pustejovsky, et al., "TimeBank 1.2" http://www.ldc.upenn.edu/, LDC, 2006.

[Saurí 05] Roser Saurí, et al., "TimeML Annotation Guidelines" http:// http://www.timeml.org/, 2005.