Applying Syntactic, Semantic and Discourse Constraints in Chinese Temporal Annotation

Nianwen Xue
Brandeis University
xuen@brandeis.edu

Yuping Zhou
Brandeis University
yzhou@brandeis.edu

Abstract

We describe a Chinese temporal annotation experiment that produced a sizable data set for the TempEval-2 evaluation campaign. We show that while we have achieved high inter-annotator agreement for simpler tasks such as identification of events and time expressions, temporal relation annotation proves to be much more challenging. We show that in order to improve the inter-annotator agreement it is important to strategically select the annotation targets, and the selection of annotation targets should be subject to syntactic, semantic and discourse constraints.

1 Introduction

Event-based temporal inference is a fundamental natural language technology that attempts to determine the temporal location of an event as well as the temporal ordering between events. It supports a wide range of natural language applications such as Information Extraction, Question Answering and Text Summarization. For some genres of text (such as news), a temporal ordering of events can be the most informative summarization of a document (Mani and Wilson, 2000; Filatova and Hovy, 2001). Temporal inference is especially important for multi-document summarization where events extracted from multiple documents need to be put in a chronological order (Lin and Hovy, 2001; Barzilay et al., 2002) to make logical sense. Event-based temporal inference is also necessary for Question Answering (Harabagiu and Bejan, 2005; Harabagiu and Bejan, 2006). For example, to answer "When was Beijing Olympics held?", events extracted from natural language text have to be associated with a temporal location, whereas to answer "how many terrorists have been caught since 9/11?", temporal ordering of multiple events is the prerequisite. Event-based temporal inference has also been studied extensively in the context of Information Extraction, which typically involves extracting unstructured information from natural language sources and putting them into a structured database for querying or other forms of information access. For event extraction, this means extracting the event participants as well as its temporal location. Generally, an event has to occur in a specific time and space, and the temporal location of an event provides the necessary context for accurately understanding that event.

Being able to infer the temporal location of an event in Chinese text has many additional applications. Besides Information Extraction, Question Answering and Text Summarization, knowing the temporal location of an event is also highly valuable to Machine Translation. To translate a language like Chinese into a language like English in which tense is grammatically marked with inflectional morphemes, an MT system will have to infer the necessary temporal information to determine the correct tense for verbs. Statistical MT systems, the currently dominant research paradigm, typically do not address this issue directly or even indirectly.

As machine learning approaches are gaining dominance in computational linguistics and producing state-of-the-art results in many areas, they have in turn fueled the demand for large quantities of human-annotated data of various types

Coling 2010: Poster Volume, pages 1363–1372,
Beijing, August 2010

that machine learning algorithms can be trained on and evaluated against. In the temporal inference domain, this has led to the creation of TimeBank (Pustejovsky et al., 2003), which is annotated based on the TimeML language (Pustejovsky et al., 2005). TimeML is becoming an ISO standard for annotating events and time expressions (ISO/TC 37/SC 4/WG 2, 2007). A version of the TimeBank has been provided as a shared public resource for TempEval-2007, the first temporal evaluation campaign aimed at automatically identifying temporal relations between events and time expressions as well the temporal ordering between events.

In this paper, we report work for a Chinese temporal annotation project as part of the 2010 multilingual temporal evaluation campaign (TempEval-2)[1]. Besides Chinese, TempEval-2 also includes English, French, Italian, Korean and Spanish. Our temporal annotation project is set up within the confines of BAT[2], a database-driven multilingual temporal annotation tool that is also used to support other TempEval-2 languages. The TempEval-2 evaluation framework takes a divide-and-conquer approach to temporal annotation. With the eventual goal being the annotation of temporal relations between events and between events and time expressions, the TempEval-2 annotation consists of a series of event and temporal annotation subtasks. The idea is that each of these subtasks will be easier to annotate than the larger task as a whole and is less demanding on the annotators. The hope is that this will lead to more consistent annotation that will be easier to learn for automatic systems as well.

The rest of the paper will be organized as follows. In Section 2, we briefly describe the seven layers of annotation. In Section 3, we describe our annotation procedure. In Section 4, we address a major issue that arises from our annotation effort, which is the question of how to select annotation targets. Our experience, some positive and some negative, shows that temporal annotation can be carried out much more smoothly and with higher quality when the right annotation targets are presented to the annotators. This is especially true during the annotation of temporal relations between events and between events and time expressions, which are more complex than simpler annotation tasks such as identifying the events and time expressions. Section 5 concludes our paper.

[1] http://www.timeml.org/tempeval2/

[2] http://www.timeml.org/site/bat

2 Layers of annotation

2.1 Events and time expressions

The ultimate goal for a temporal annotation project is to determine the temporal relationship between events, and between events and time expressions. In order to achieve that objective, events and time expressions must be first identified. Specifically, this means marking up text spans in a document that can be used to represent the events and time expressions. Events in particular are abstract objects and a full description of an event would include its participants and temporal and spatial location. The TempEval annotation framework simplifies this by just marking a verb or a noun that best represents an event. The verb or noun can be considered as an "event anchor" that represents the most important aspect of the event. This is illustrated in (1), where the verbs 参加 ("attend"), 举行 ("hold") and the noun 仪式 ("ceremony") are marked as event anchors.

(1) 国务院 副总理 邹家华
State Council Vice Premier Zou Jiahua
参加 了 今天 举行 的 投产
attend ASP today hold DE commissioning
剪彩 仪式 。
ribbon-cutting ceremony .

"Vice Premier Zou Jiahua of the State Council attended today's commissioning ribbon-cutting ceremony".

Once the text spans of event anchors are annotated, these events are then annotated with a set of attributes. The TempEval annotation framework allows variations across languages in the number of attributes one can define as well as the values for these attributes. For example, in the English annotation, one of the event attributes is grammatical *tense* which can be read off the morphological inflections of a verb. Chinese verbs, on the other hand, are not inflected for tense. Instead, in the

Chinese annotation, we have a more fully developed *aspect* attribute that has eight possible values: *Actual*, *Experiential*, *Complementive*, *Delimitative*, *Progressive*, *Durative*, *Inceptive*, and *Continuative*, largely based on the theoretical work of Xiao and McEnery (2004).

The most important attribute for both English and Chinese, however, is the *Class* attribute. The values for this attribute include *Reporting*, *Aspectual*, *Perception*, *I-Action*, *I-State*, *State*, and *Occurrence*. The different values of the *Class* attribute effectively constitute a classification of events, and they are defined in the TimeML specification language (Pustejovsky et al., 2005).

The other building block in the TempEval annotation framework is time expressions. Like events, time expressions are marked with both text spans and a set of attributes. The annotation of time expressions is relatively straightforward, and we follow the TimeML standards in our annotation study. In TimeML, time expressions are formally called TIMEX3s, and they have two obligatory attributes: *Type* and *Value*. The value of *Type* is one of *time*, *date*, *duration* or *set*. The *Value* attribute is essentially a normalized time value based on the TIDES standard for annotating time expressions (Ferro et al., 2004). The normalization allows easy comparison of time expression. For example, there are three time expressions in (2), 一九九二年("1992"), 一九九六年 ("1996") and 今年 ("this year"). Note that even though 一九九二年 至 一九九六年 (" 1992 to 1996") forms one duration, it is annotated as two time expressions. All three time expressions in the sentence are dates, and their normalized values are *1992*, *1996*, and *1997* respectively. To determine the normalized value for 今年 ("this year"), we need to know the document creation time, and fortunately this information is available in the metadata for the Chinese Treebank documents.

(2) 一九九二年 至 一九九六年 上海
1992 to 1996 Shanghai
国内 生产 总值 年均
GDP per year on average
增长 百分之十四点二 ， 今年 的
grow 14.2% , this year DE
增长 速度 也 将 达到 百分之十三
growth speed also will reach 13%
以上 。
above

"From 1992 to 1996, Shanghai's GDP on average grows at14.2% per year. This year the (GDP) growth will also reach above 13%."

2.2 Temporal relations

Once the events and time expressions are in place, we are in a position to annotate various temporal relations that are defined over them. (Since events and time expressions are entities that temporal relation is defined upon, we will subsume them under the cover term "temporal entity" when convenient.) The ultimate goal of temporal annotation is to identify all temporal relations in text. This goal cannot be achieved by manually annotating temporal relation of all temporal entities for three reasons. First, it is infeasible, given the number of temporal entities in a typical document. Second, it is unnecessary due to the transitive property of certain types of temporal relation. For example, if *e1*, *e2* and *e3* are all events, and if *e1* is before *e2*, and *e2* is before *e3*, there is no need to also annotate the relation between *e1* and *e3*. Third, the result of annotating all temporal entity pairs does not reflect the natural temporal relations that exist in text. Verhagen et al. (2009) found that a major contributor to high inter-annotator disagreement was hard-to-classify cases that annotators were instructed not to avoid. If a temporal relation is not made clear in text, then it should not be present in annotation.

Since it is infeasible, unnecessary and even detrimental to manually annotate all possible relations between temporal entities, the question then becomes one of selecting which temporal relations to annotate. The TempEval-2 evaluation starts by annotating the following temporal relations, which it considers to be a priority:

1. between an event and a time expression
2. between an event and the document creation time
3. between a subordinating event and its corresponding subordinated event

4. between a main event and its immediately preceding main event

The TempEval-2 annotation uses six values for all temporal relations, and they are *Before*, *Before-or-Overlap*, *Overlap*, *Overlap-or-After*, *After* and *Vague*. The *Vague* value is only used as the last resort when the annotator really cannot determine the temporal relationship between a pair of temporal entities. In the meantime, the TempEval-2 also allows variations from language to language regarding specific annotation strategies for each subtask. For Chinese temporal annotation, most of the decisions we have to make revolve around one central question, and that is which temporal entity pair to annotate.

2.2.1 Relation between events and time expressions

The annotation of the relationship between events and time expressions involves i) determining which event is related to which time expression, and ii) what is the nature of this relationship. In (3), for example, there are three events and three time expressions that enter into the temporal relation annotation. If the annotator is required to annotate all possible event/time combinations, there will be nine possible pairs. There are at least three possible strategies to go about selecting event/time pairs to annotate. The first strategy is to annotate all possible pairs. This seems to add unnecessary burden to the annotator because if we know that *e1* overlaps *t1*, we can infer the temporal relationship between *e1* and *t3* by virtue of the fact that *t1* occurs before *t3*. The second strategy is to allow the annotator to freely choose which event/time pair to annotate based on whether there is a clear temporal relation between them. This eliminates the possibility that the annotator is forced to annotate hard-to-classify and inconsequential relations, but leaving this decision to the annotator entirely might lead to low inter-annotator agreement where annotators choose to annotate different event/time pairs.

(3) 国际 货币 基金 组织 [**t1** ２１日]
International Monetary Fund 21st
在 此间 [**e1** 发表] 一 份 临时
at here publish one CL preliminary
评估 报告 ， 再次 [**e2** 调低] 了
assessment report , again lower AS
它 对 [**t2** 今] [**t3** 明] 两 年
its regarding this next two year
全球 经济 增长 速度 的 [**e3**
global economic growth speed DE
预测] 。
forecast .

"The International Monetary Fund on 21 published a preliminary assessment report, again lowering its forecast of the global economic growth for this year and next year."

In our annotation, we adopt a third strategy. Instead of simply asking which event bears a temporal relation to which temporal expression in the same sentence, we ask annotators to judge *which event(s) a given temporal expression is **intended** to modify*. In essence, this amounts to asking the annotator to first make a syntactic decision about which events fall within the scope of a time expression. In (3), all three events *e1*, *e2* and *e3* fall within the scope of *t1*, and none of them are in the scope of *t2* and *t3*. This approach reduces the number of fuzzy temporal relations that annotators might disagree on due to preference for thoroughness vs. accuracy.

2.2.2 Temporal relation between subordinating event and subordinated event

The two tasks in the TempEval framework that deal with event pairs are to annotate temporal relation between the subordinating event and the subordinated event, as well as the relation in main event pairs. The division of labor between them is quite clear: the former deals with intra-sentential temporal relations whereas the latter handles inter-sentential relations. It is not immediately clear, however, how each of the two types of relations should be defined.

Unlike in the event/time annotation where syntactic notions are invoked in selecting event/time pairs to annotate, our definitions of subordinating and subordinated events are primarily based on semantic criteria. The subordinating event is roughly the predicate while the subordinated event is one of its arguments, provided that both the

predicate and the argument are anchors of events. For example, in (4), there are two subordinating and subordinated event pairs. *e2* is a subordinated event of *e1*, and *e4* is a subordinated event of *e3*.

(4) 广东 [**e1** 举行] [**e2** 研讨会] [**e3** 介绍] [**e4** 税改] 及 加工 贸易 台帐 制度
Guangdong hold symposium introduce tax reform and processing trade accounting regulation

"Guangdong held a symposium introducing the tax reform and the accounting regulations on processing trade."

An alternative to using the notion of predicate-argument structure in determining the subordinating/subordinated events is to resort to syntactic relations such as the verb and its object. The net result would be the same for Example (4). However, the same argument that motivates the annotation of the predicate-argument structures in the Propbank (Palmer et al., 2005) and the Chinese Propbank (Xue and Palmer, 2009) also applies to temporal annotation. That is, the predicate-argument structure and temporal relations tend to hold constant in spite of the syntactic alternations and variations. For example, the temporal relation between the noun 研讨会 ("symposium") event and the verb 举行 ("hold") event remains the same in (5) in spite of the change in the syntactic relation between them. If only event pairs in a verb-object relation are annotated, the temporal relation between *e2* and *e1* in (5) would be lost.

(5) [**e2** 研讨会] 在 广东 [**e1** 举行]
symposium PREP Guangdong hold

"The symposium was held in Guangdong."

2.2.3 Temporal relations between main events

The purpose of annotating the temporal relation between main events is to capture the temporal ordering of events scattered in different sentences that constitute the main chain of events covered in the article. Annotation of the temporal relation between main events is further divided into two steps. In the first step, main events are first identified among all events in a sentence, and then the temporal relation between the main events in adjacent pairs of sentences is annotated. As a first approximation, we define "main event" as follows: a main event is the event expressed by the main verb of the top-most level clause of a sentence. The underlying assumption is that good writing would place words representing important events in prominent positions of a sentence and the first choice of a prominent position in a sentence is probably the main verb. An additional stipulation is that in case of a co-ordinated construction involving two or more main verbs at the top-most level, the event represented by the first is the main event of the sentence. This is to ensure that each sentence has only one main event. As we shall see in Section 3, this seemingly simple turns out to be surprisingly difficult, as reflected in the low inter-annotator agreement.

2.2.4 Temporal relation between events and the document creation time

In this layer, all the events identified in a document are annotated according to their temporal relation to the document creation time. This task is particularly challenging and intellectually interesting for Chinese. As an isolating language (Li and Thompson, 1981), Chinese has a small word to morpheme ratio. That is, the majority of its words consist of single morphemes. As a result, it lacks the inflectional morphology that grammatically marks tense. Tense directly encodes the temporal location of an event in natural language text and the lack of observable grammatical tense makes it that much harder to determine the temporal location of an event in Chinese text. This is not to say, however, that Chinese speakers do not attempt to convey the temporal location of events when they speak or write, or that they cannot interpret the temporal location when they read Chinese text, or even that they have a different way of representing the temporal location of events. In fact, there is evidence that the temporal location is represented in Chinese in exactly the same way as it is represented in English and most world languages: in relation to the moment of speech. One piece of evidence to support this claim is that Chinese temporal expressions like 今天 ("today"), 明天 ("tomorrow") and 昨天 ("yesterday") all assume a

temporal deixis that is the moment of speech in relation to which all temporal locations are defined. Annotating the temporal relation between events and document creation time would then directly capture the temporal location of events.

3 Annotation procedure and annotation consistency

The data set consists of 60 files taken from the Chinese Treebank (Xue et al., 2005). The source of these files is Xinhua newswire. It goes through a two-phase double blind and adjudication process. The first phase involves three annotators, with each file annotated by two annotators; the second phase involves two judges, with each double annotated document assigned to a single judge for disagreement resolution. The inter-annotator agreement between the two annotators (A and B) as the agreement between each annotator and the judge (J) are presented in Table 1. The agreement is measured in terms of F1-score[3], which is a weighted average between precision and recall. The F1-score is calculated as follows:

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (1)$$

The agreement statistics in Table 1 clearly show that event and time expression annotations are easier but temporal relations are harder as reflected in the lower inter-annotator agreement scores. This is somewhat expected because relations involve two temporal entities while we are only dealing with one temporal entity with event and time expression annotations. The figures also show the seemingly simple task of main event annotation (which only involves picking one event per sentence as the main event) has a surprisingly low inter-annotator agreement score. One reason might be that in a less grammaticalized language like Chinese, it is not always clear which verb is the main verb when the syntactic tree information is not displayed in the annotation interface. Another reason is that annotators sometimes disregard the syntax-based rule when it runs too much afoul to their intuition, a point that we will come back to and discuss in greater detail in Section 4.

[3]For a subset of the tasks, the total number of annotated instances for the two annotators is the same. This subset includes identification of main events, the temporal relation between the main events in two adjacent sentences, and the temporal relation between an event and the document creation time.

Layer	f(A, B)	f(A, J)	f(B, J)
event-extent	0.90	0.93	0.94
timex-extent	0.86	0.88	0.93
main-events	0.74	0.90	0.82
tlinks-main-events	0.65	0.70	0.75
tlinks-dct-events	0.77	0.86	0.90
tlinks-e-t	0.75	0.88	0.83
tlinks-sub-e	0.53	0.74	0.70

Table 1: Inter-annotator agreement for the subtasks: *event-extent*, the textual extent of an event anchor; *timex-extent*, the textual span of a time expression; *tlinks-main-event*, the temporal relation between the main events; *tlinks-dct-events*, the temporal link between an event and the document creation time; *tlinks-e-t*, the temporal relation between an event and a time expression; *tlinks-sub-e*, the temporal relation between a subordinating event and a subordinated event.

It is worth noting that the annotation of the temporal relation between an event and a time expression, and between a subordinating event and a subordinated event involves two decisions. The annotator needs to first decide which pairs of temporal entities to annotate, and then decide what temporal relation should be assigned to each temporal entity pair. To take a closer look at which of these two decisions creates more of a problem for the annotator, we computed the agreement figures for these two steps respectively. In Table 2, Column 3 presents the figure for just identifying which pair to annotate, and Column 4 is the agreement for just assigning the temporal relation, assuming the same pair of temporal entities are found by both annotators.

Layer	all	identification f	relation
tlinks-e-t	0.75	0.86	0.89
tlinks-sub-e	0.53	0.60	0.87

Table 2: Detailed agreement for event-time and subordinating-subordinated events

From Table 2, it is clear that for both tasks,

there is lower agreement between the annotators in deciding which pair to annotate. Once the two annotators agree on which pair to annotate, determining the temporal relation is relatively easier, as reflected in higher agreement.

4 Detailed discussion

As described in Section 2, when annotating the temporal relation between an event and a time expression, the annotators are instructed to annotate an event-time pair if the event is falling within the syntactic scope of the time expression. When annotating the relation between subordinating and subordinated events, the annotators are instructed to select event pairs based on the semantic notion of predicate-argument structure. This assumes a certain level of linguistic sophistication on the part of the annotators. From the lower agreement score in identifying event-time pairs (Table 2), it is clear that our annotators, who are not trained linguists, lack in this type of specialized knowledge. They are better at making the more intuitive judgment regarding the temporal relation between two temporal entities. One solution is obviously to find better trained linguists to perform these tasks, but it may not always be feasible. Since our data is taking from the Chinese Treebank and has already been annotated with syntactic structures and predicate-argument structures (from the Chinese Propbank annotation (Xue and Palmer, 2009)), an alternative is to extract the event-time or event-event pairs using the syntactic and predicate-argument structures as constraints.[4]

The annotation of main events and their relations presents a different challenge. Our first approximation is to select main events based on syntactic considerations. A main event is equated with the matrix verb in a sentence. In many cases this turns out to be unintuitive. Two of the recurring counter-intuitive cases involve directly quoted speech and coordination structures.

Directly quoted speech In Chinese newswire text, it is often the case that the source of information is explicitly cited in the form of direct quotations. (6) is such an example:

(6) 宋健 说：“如今 ，中国
Song-Jian say , “ nowadays , China
已 能 生产 上万 门
already can produce tens-of-thousands CL
数字 电话 程控交换机。”
digital telephone PBX

“Song Jian said, ‘nowadays, China is capable of producing tens of thousands of digital telephone PBX.’ ”

While the event represented by the underlined verb 说 (“say”) may very well be important in some natural language processing applications (for example, sometimes the source of the target information is crucial), it is not normally part of the intended information being covered by a news article. And it does not make much sense to annotate its temporal relation to adjacent main events that are on a par with what was said, not the saying event itself. The point would be even clearer when such a case is contrasted with a case in which a similar semantic relation is formulated in a different syntactic structure, as shown in (7):

(7) **据** 官方 权威 人士
according to official authority source
透露 ，今年 中国 政府
divulge , this-year China government
确定 的 经济 增长率 为
determine DE economic growth rate be
百分之八。
8%

“According to some official sources in position of authority, the economic growth rate determined by the Chinese government is 8%.”

Because of the presence of the preposition 据 (“according to”), the underlined reporting verb 透露 (“divulge”), similar to 说 (“say”) in (6) with respect to its semantic relation to the following material, would not be annotated as representing the main event of the sentence. The difference in the annotation of the main event between (7) and (6) seems to be an undesirable artifact of the purely syntax-based annotation rule for identifying main events.

[4] See a similar approach in Bethard et al. (2007).

Co-ordination structure Co-ordination by no means is a rare occurrence in the data, and often times, all events within a co-ordination structure, taken together, represent the main event of the sentence. For example, in (8), both events represented by the underlined verbs seem to be equally significant and should be included in the same chain of events. Given the prevalence of co-ordination between verbs, the stipulation that only the first one counts significantly undermines the coverage of the task and goes against the annotator's intuitions.

(8) 今年 9月 ，多 家 外国
This year September , many CL foreign
石油公司 与 哈 国家 石油
oil company with Kazakstan national oil
公司 签署 了 一揽子 “世纪
company sign LE a series of “century
合同” ，这些 合同 将 在今后
contract” , these contract will in future
４０年 内 产生 ７０００亿
40 years within generate 700-billion
美元 的 巨额 利润 。
dollar DE enormous profit

“In September of this year, many foreign oil companies signed a series of ‘century contract’ with Kazakstan National Oil Company. These contracts will generate an enormous profit of 700-billion dollars.”

The issue in the annotation of the temporal relation between main events seem to be more in the selection of main event pairs than in the determination of the nature of their relationship. Our current rule states that any two main events in consecutive sentences form a pair for annotation. This task suffers a low level of inter-annotator agreement partly because many main events identified by syntactic criteria are not actually main events in our intended sense. Often times, two consecutive main events come from different levels of the discourse structure or different chains of events, which puts annotators in a hard-to-classify situation.

To achieve high inter-annotator consistency when annotating the temporal relation between events from different sentences, we believe the selection of event pairs has to be informed by the discourse structure of the document. This only makes sense given that the annotation of temporal relation between events and time expressions within one sentence is informed by the syntactic structure, and the temporal relation between subordination and subordinating events benefits from an understanding of the predicate-argument structure.

The specific type of discourse structure we have in mind is the kind represented in the Penn Discourse Treebank (Miltsakaki et al., 2004). The Penn Discourse Treebank-style of annotation can inform temporal relation annotation in at least two ways. First, the Penn Discourse Treebank annotates the discourse relation between two adjacent sentences. The discourse relation holds between two abstract objects such as events or propositions. If a discourse relation holds between two events, the temporal relation between those two events might also be what we are interested in for temporal annotation. The implicit assumption is that the discourse structure of a document represents the important temporal relations within that document as well. (9) is an example taken from the Penn Discourse Treebank. The discourse relation, characterized by the discourse connective “in particular”, holds between the events anchored by“dropped” and “fell”. The temporal relation between these events also happens to be what we would be interested in if we are to annotate the main events between two adjacent sentences. Notice that in (9), material that is irrelevant to the discourse relation is taken out of the two arguments of this discourse relation, which are marked in italics and bold face respectively.

(9) *Meanwhile, the average yield on taxable funds dropped nearly a tenth of a percentage point, the largest drop since midsummer.* implicit = in particular **The average seven-day compound yield**, which assumes that dividends are reinvested and that current rates continue for a year, **fell to 8.47%, its lowest since late last year, from 8.55% the week before, according to Donoghue’ s.**

The Penn Discourse Treebank also marks attributions when annotating discourse relations. In

(10), for example, "he says" will be marked as a case of attribution and the "say" verb would be marked as the main event of the sentence if syntactic criteria are followed. Having attributions identified would directly help with the temporal annotation of examples like (6), where the main event is embedded in direct quoted speech.

(10) When *Mr. Green won a $240,000 verdict in a land condemnation case against the State in June 1983*, [he says] **Judge O' Kicki unexpectedly awarded him an additional $100,000.**

As of now, the data we use for our temporal annotation experiment have not yet been annotated with discourse structures. In order to make our temporal annotation sensitive to the discourse structure, we either have to annotate the discourse structure in a separate pass, or to incorporate the key elements of the discourse structure when developing guidelines for temporal annotation.

5 Conclusion

We described a Chinese temporal annotation experiment that produced a sizable data set for the TempEval-2 annotation campaign. We show that while we have achieved high inter-annotator agreement for simpler tasks such as identification of events and time expressions, temporal relation annotation proves to be much more challenging. We show that in order to improve annotation consistency it is important to strategically select the annotation targets, and this selection process should be subject to syntactic, semantic and discourse constraints.

Acknowledgements

This work is supported by the National Science Foundation via Grant No. 0855184 entitled "Building a community resource for temporal inference in Chinese". All views expressed in this paper are those of the authors and do not necessarily represent the view of the National Science Foundation.

References

Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

Steven Bethard, James H. Martin, and Sara Klingenstein. 2007. Finding Temporal Structure in Text: Machine Learning of Syntactic Temporal Relations. *International Journal of Semantic Computing*, 11(4).

Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth Sundheim, and George Wilson. 2004. TIDES 2003 Standard for the Annotation of Temporal Expressions.

Elena Filatova and Eduard Hovy. 2001. Assigning Time-Stamps to Event Clauses. In *Proceedings of the Workshop on Temporal and Spatial Information Processing*, Toulouse.

Sanda Harabagiu and Cosmin Adrian Bejan. 2005. Question Answering Based on Temporal Inference. In *Proceedings of the AAAI-2005 Workshop on Inference for Textual Question Answering*, Pittsburgh, Pennsylvania.

Sanda Harabagiu and Cosmin Adrian Bejan. 2006. An Answer Bank for Temporal Inference. In *Proceedings of LREC 2006*, Genoa, Italy.

ISO/TC 37/SC 4/WG 2. 2007. Language Resource Management – Semantic Annotation Framework (SemAF) – Part 1: Time and events.

Charles Li and Sandra Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. Berkeley, Los Angeles, London: University of California Press.

Chin-Yew Lin and Eduard Hovy. 2001. Neats: A multidocument summarizer. In *Proceedings of the Document Understanding Workshop*.

Inderjeet Mani and George Wilson. 2000. Robust temporal processing of news. In *Proceedings of the ACL'2000*, Hong Kong, China.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse TreeBank. In *Proceedings of the Language Resources and Evaluation Conference*, Lisbon, Portugal.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, David Day, Lisa Ferro, Robert Gaizauskas, Marcia Lazo, Andrea Setzer, and Beth

Sundheim. 2003. The TimeBank Corpus. *Corpus Linguistics*, pages 647–656.

James Pustejovsky, Bob Ingria, Roser Sauri, Jose Castano, Jessica Littman, Rob Gaizauskas, Andrea Setzer, G. Katz, and I. Mani. 2005. The specification language TimeML. In I. Mani, J. Pustejovsky, and R. Gaizauskas, editors, *The Language of Time: a Reader*. Oxford University Press.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. 2009. The TempEval Challenge: Identifying Temporal Relation in Text. *Language Resources and Evaluation*, 43(1):161–179.

Richard Xiao and Tony McEnery. 2004. *Aspect in Mandarin Chinese: A Corpus-based Study*. Amsterdam: John Benjamins.

Nianwen Xue and Martha Palmer. 2009. Adding semantic roles to the Chinese Treebank. *Natural Language Engineering*, 15(1):143–172.

Nianwen Xue, Fei Xia, Fu dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2):207–238.