Discourse-constrained Temporal Annotation

Yuping Zhou
Brandeis University
Waltham, MA 02452
yzhou@brandeis.edu

Nianwen Xue
Brandeis University
Waltham, MA 02452
xuen@brandeis.edu

Abstract

We describe an experiment on a temporal ordering task in this paper. We show that by selecting event pairs based on discourse structure and by modifying the pre-existent temporal classification scheme to fit the data better, we significantly improve inter-annotator agreement, as well as broaden the coverage of the task. We also present analysis of the current temporal classification scheme and propose ways to improve it in future work.

1 Introduction

Event-based temporal inference is a fundamental natural language technology aimed at determining the temporal anchoring and relative temporal ordering between events in text. It supports a wide range of natural language applications such as Information Extraction (Ji, 2010), Question Answering (Harabagiu and Bejan, 2005; Harabagiu and Bejan, 2006) and Text Summarization (Lin and Hovy, 2001; Barzilay et al., 2002). Creating consistently annotated domain-independent data sufficient to train automatic systems has been the bottleneck. While low-level temporal annotation tasks such as identifying events and time expressions are relatively straightforward and can be done with high consistency, high-level tasks necessary to eventually arrange events in a document in a temporal order have proved to be much more challenging.

Among these high-level tasks, the task of annotating the temporal relation between main events stands out as probably the most challenging. This task was the only task in the TempEval campaigns (Verhagen et al., 2009; Verhagen et al., 2010) to deal with inter-sentential temporal relations, and also the only one to directly tackle event ordering. The idea is that events covered in an article are scattered in different sentences, with some, presumably important ones, expressed as predicates in prominent positions of a sentence (i.e. the "main event" of the sentence). By relating main events from different sentences of an article temporally, one could get something of a chain of important events from the article.

This task, in both previously reported attempts, one for English (Verhagen et al., 2009) and the other for Chinese (Xue and Zhou, 2010), has the lowest inter-annotator agreement (at 65%) among all tasks focusing on annotating temporal relations. Verhagen et al. (2009) attribute the difficulty, shared by all tasks annotating temporal relations, mainly to two factors: rampant temporal vagueness in natural language and the fact that annotators are not allowed to skip hard-to-classify cases.

Xue and Zhou (2010) take a closer look at this task specifically. They report that part of the difficulty comes from "wrong" main events (in the sense that they are not main events in the *intended* sense) being selected in the preparation step. This step is a separate task upstream of the temporal relation task. The "wrong" main events produced in this step become part of event pairs whose temporal relation it makes no sense to annotate, and often is hard-to-classify. The reason "wrong" main events get selected is because the selection is based on syntactic criteria. In fact, these syntactic criteria produce results so counter-intuitive that this seemingly simple

Proceedings of the Fifth Law Workshop (LAW V), pages 161–169,
Portland, Oregon, 23-24 June 2011.

preparation task only achieves 74% inter-annotator agreement.

Another part of the difficulty comes from mechanical pairing of main events for temporal relation annotation. Simply pairing up main events from adjacent sentences oversimplifies the structure within an article and is prone to produce hard-to-classify cases for temporal relation annotation. Both causes point to the need for a deeper level of text analysis to inform temporal annotation. For this, Xue and Zhou (2010) suggest introduction of discourse structure as annotated in the Penn Discourse Treebank (PDTB) into temporal relation annotation.

So the previous two reports, taken together, seem to suggest that the reason this task is especially challenging is because the difficulty associated with temporal vagueness in natural language, which is shared by all tasks dealing with temporal relation, is compounded by the problem of having to annotate farfetched pairs that should not be annotated, which is unique for the only task dealing with inter-sentential temporal relations. These two problems are the foci of our experiment done on Chinese data.

The paper is organized as follows: In Section 2, we describe the annotation scheme; in Section 3, we describe the annotation procedure; in Section 4 we report and discuss the experiment results. And finally we conclude the paper.

2 Annotation Scheme

As stated in the introduction, there are two problems to be addressed in our experiment. The first problem is that "wrong" main events get identified and main events that do not bear any relation are paired up for temporal annotation. To address this problem, we follow the suggestion by Xue and Zhou (2010), namely using a PDTB-style discourse structure to pick out and pair up main events. We believe that adopting a discourse-constrained approach to temporal annotation will not only improve annotation consistency but also increase the *Informativeness Value* of the annotated data, under the assumption that temporal relations that accord with the discourse structure are more valuable in conveying the overall information of a document. Since there is no Chinese data annotated with PDTB-style discourse structure available, we have to develop our own. The scheme for this step is described in Section 2.1.

The second problem is that there is too much temporal vagueness in natural language with respect to the temporal classification scheme. Since we cannot change the way natural language works, we try to model the classification scheme after the data it is supposed to classify. The scheme for the temporal annotation is covered in Sections 2.2 and 2.3.

2.1 Discourse-constrained selection of main events and their pairs

2.1.1 Discourse annotation scheme

The PDTB adopts a lexically grounded approach to discourse relation annotation (Prasad et al., 2008). Based on discourse connectives like "*since*", "*and*", and "*however*", discourse relation is treated as a predicate taking two *abstract objects* (AO's) (such as events, states, and propositions) as arguments. For example, in the sentence below, "*since*" is the lexical anchor of the relation between Arg1 and Arg2 (example from Prasad et al. (2007)).

(1) Since [$_{Arg2}$ McDonald's menu prices rose this year], [$_{Arg1}$ the actual decline may have been more].

This notion is generalized to cover discourse relations that do not have a lexical anchor, i.e. implicit discourse relations. For example, in the two-sentence sequence below, although no discourse connective is present, a discourse relation similar to the one in (1) is present between Arg1 and Arg2 (example from Prasad et al. (2007)).

(2) [$_{Arg1}$ Some have raised their cash positions to record levels]. [$_{Arg2}$ High cash positions help buffer a fund when the market falls].

Based on this insight, we have fashioned a scheme tailored to linguistic characteristics of Chinese text. The linguistic characteristics of Chinese text relevant to discussion here can be illustrated with the following sentence.

(3) 据悉 , [$_{AO1}$ 东莞 海关
according to reports , Dongguan Customs
共 接受$_{e1}$ 企业 合同 备案
in total accept company contract record
八千四百多 份] , [$_{AO2}$ 比 试点 前
8400 plus CL , compare pilot before

略 有$_{e2}$ 上升] ，[$_{AO3}$企业
slight EXIST increase , company
反应$_{e3}$ 良好] ，[$_{AO4}$普遍
respond/response well/good , generally
表示$_{e4}$ 接受] 。
acknowledge accept/acceptance

"According to reports, [$_{AO1}$ Dongguan District Customs accepted$_{e1}$ more than 8400 records of company contracts], [$_{AO2}$ (showing$_{e2}$) a slight increase from before the pilot]. [$_{AO3}$ Companies responded$_{e3}$ well], [$_{AO4}$ generally acknowledging$_{e4}$ acceptance]."

One feature is that it is customary to have complex ideas packed into one sentence in Chinese. The sentence above reports on how a pilot program worked in Dongguan City. Because all that is said is about the pilot program, it is perfectly natural to include it all in a single sentence in Chinese. Intuitively though, there are two different aspects of how the pilot program worked: the number of records and the response from the affected companies. To report the same facts in English, it is probably more natural to break them down into two sentences, but in Chinese, *not only are they merely separated by comma, but also there is no connective relating them*.

Another feature is that grammatical relation between comma-separated chunks within a sentence is not always clear. In the above sentence, for instance, although the grammatical relations between AO1 and AO2, and between AO3 and AO4 are clear in the English translation (i.e. the first in each pair is the main clause and the second an adjunct), it is not at all clear in the original. This is the result of several characteristics of Chinese, for example, there is no inflectional clues on the verb to indicate its grammatical function in the sentence.

Based on these features of Chinese text[1], we have decided to use punctuation as the main potential indicator for discourse relations: the annotator is asked to judge, at every instance of comma, period, colon and semi-colon, if it is an indicator for discourse relation; if both chunks separated by the punctuation are projections of a predicate, then there is a discourse relation between them. Applying this scheme to the sentence in (3), we have four abstract objects as marked up in the example.

[1] A more detailed justification for this scheme is presented in Zhou and Xue (2011).

To determine the exact text span of each argument of a relation, we adopt the *Minimality Principle* formulated in Prasad et al. (2007): only as many clauses and/or sentences should be included in an argument selection as are minimally required and sufficient for the interpretation of the relation. Applying this principle to the sentence in (3), we can delimit the three sets of discourse relations as follows: AO1–AO2, (AO1,AO2)–(AO3,AO4), and AO3–AO4.

2.1.2 Selection and pairing-up of main events

Selection of main events is done on the level of the *simplex* abstract object, with one main event per simplex AO. The main event corresponds to the predicate heading the simplex AO. In (3), there are four simplex AO's, AO1-4 (which further form two *complex* AO's, (AO1,AO2) and (AO3,AO4)). The anchors for the four main events are the underlined verbs labeled as "*e1-4*".

Pairing up the main events is done on the level of discourse relation. In the case of a relation only involving simplex AO's, the main events of the two AO's pair up; in the case of a relation involving complex AO's, the discourse relation is distributed among the simplex AO's to form main event pairs. For example, with the discourse relation (AO1,AO2)–(AO3,AO4), four pairs of main events are formed: *e1–e3*, *e1–e4*, *e2–e3*, and *e2–e4*. This gets tedious fast as the number of simplex AO's in a complex AO increases; in this experiment, the annotator relies on her discretion in such cases. This problem should be addressed in a more elegant way in the future.

It is worth noting that in addition to picking out right main events and event pairs for temporal annotation, this scheme also broadens the coverage of the task. In the old scheme based on syntactic criteria, there is a stipulation: one main event per sentence. Because the new discourse-constrained scheme is tailored to the characteristics of Chinese text, it is able to expose more main events (in the intended sense) to temporal annotation.

2.2 Classification scheme for temporal relation annotation

By modifying the six-value scheme used in TempEval (containing *before, overlap, after, before-or-overlap, overlap-or-after* and *vague*), our classifica-

tion scheme has seven values in it: *before, overlap, after,* ***not-before***, ***not-after***, ***groupie***, and ***irrelevant***.

2.2.1 The values "*not-before*" and "*not-after*"

The values "*not-before*" and "*not-after*" are equivalent to "*overlap-or-after*" and "*before-or-overlap*" in the TempEval scheme. The reason we made this seemingly vacuous change is because we found that the old values were used for two different purposes by annotators. In addition to their intended use, i.e. to capture indeterminacy between the two simplex values, they were also used to label a specific case of "*overlap*". An example of such misuse of the value "*before-or-overlap*" is presented below:

(4) 一九九六 年 ， [$_{e1}$ 产生] 了 第一 位
1996 year , generate ASP first CL
本地 华人 法官 ， 到 目前 ，
local Chinese judge , until at present ,
已 有 近 二十 位 本地 华人 [$_{e2}$
already EXIST close 20 CL local Chinese
担任] 司法 官员 。
hold the post judicial official .

"The first local ethnic Chinese judge [$_{e1}$ assumed] the office in 1996; up until now, there have been close to 20 ethnic Chinese locals [$_{e2}$ holding] the posts of judicial officials."

The reason for such use is probably because it represents two alternative ways of looking at the temporal relation between the two events : either *e1* is *before* the later bulk of *e2* or *e1 overlaps* the beginning tip of *e2*. To avoid such mis-uses, we made the above change.

2.2.2 The value "*groupie*"

This value is set up for two events whose temporal relation to each other is unclear, but are known to happen within the same temporal range. For example, the temporal relation between the events represented by the underlined verbs should be classified as "*groupie*".

(5) 今 昨 两 天 ， 香港 特区
today yesterday two day , Hong Kong SAR
全国政协 委员 还 [$_{e1}$ 视察] 了 宁波
CPPCC member also inspect ASP Ningbo
开发区 、 宁波 西田信 染织
development district , Ningbo Xitianxin Textile
有限公司 ， [$_{e2}$ 游览] 了 天一阁 、
Ltd. , tour ASP Tianyi Pavilion ,
蒋氏 祖居 。
Chiang ancestral home .

"Yesterday and today, CPPCC members from Hong Kong SAR also [$_{e1}$ visited] Ningbo Development District and Ningbo Xitianxin Textile Ltd., and [$_{e2}$ toured] Tianyi Pavilion and the ancestral home of Chiang Kai-shek."

In this example, the common range shared by the two events is expressed in the form of a time expression, "今昨两天" ("yesterday and today"), but it does not have to be the case. It can be in the form of another event (e.g., "工程建设过程中" ("during the process of project construction")), or another entity with a time stamp (e.g., "八五期间" ("in the Eighth Five-year Plan period")).

It should be noted that the linguistic phenomenon captured by this value can occur in a situation where the internal temporal relation between two events can be classified with another value. So ideally, this value should be set up as a feature parallel to the existent classification scheme. But due to technical restrictions imposed on our experiment, we grouped it with all the others and instructed the annotators to use it only when none of the five more specific values applies.

2.2.3 The value "*irrelevant*"

We substituted this value for the old one "*vague*" because it is too vague. Anything that cannot fit into the classification scheme would be labeled "*vague*", but in fact, some cases are temporally relevant and probably should be characterized in the classification scheme. Case in point are those we now label "*groupie*".

This change reflects our guiding principle for designing the classification scheme. If the relation between two events is temporally relevant, we should try to characterize it in some way; if too many relations are temporally relevant but too vague to fit into the classification scheme (comfortably), then the adequacy of the scheme is questionable.

2.3 An additional specification: which event?

In addition to the classification scheme, it is also necessary to specify which event should be considered for temporal annotation. This question has

never been clearly addressed, probably because it seems self-evident: the event in question is the one expressed by the event anchor (usually a verb). This intuitive answer actually accounts for some too-vague-to-classify cases. In some cases, the event that is easily annotated (and should be the one being annotated in our opinion) is not the event expressed by the verb, as is the case in (6).

(6) 在 吸收 外商 投资 方面 ,
PREP absorb foreign business invest aspect ,
中国 现 已 成为 世界 上 利用
China now already become world POSTP utilize
外资 最多 的 发展中 国家。
foreign fund most DE developing country.

"With regard to attracting foreign business investments, China has now become the developing country that utilizes the most foreign funds in the world."

This sentence is taken from an article summarizing China's economic progress during the "Eighth Five-Year Plan" period (from 1991 to 1995). The anchor for the main event of the sentence is clearly "成为" ("become"), but should the event it represents, the process of China becoming the developing country that utilizes the most foreign funds, be considered for the temporal relation annotation? It is both counter-intuitive and impractical.

Intuitively, the sentence is a statement of the *current* state with regard to attracting foreign business investments, not of the process leading up to that state. If we were to consider the process of "becoming" in relation to other events temporally, we would have to ask, *when are the starting and ending points of this process?* How does one decide when it is not made clear in the article? One could conceivably go as far back as to when China did not use one cent of foreign funds. Should it be restricted to the "Eighth Five-Year Plan" period since it is the target period of the whole article? But why use the five-year period, when there are more specific, syntactically explicit aspectual/temporal modifiers in the sentence, i.e. "现已" ("now already"), to restrict it? To make use of these in-sentence aspectual/temporal modifiers, we have to go with our intuition that the event is the current state of China with regard to utilizing foreign investments, i.e. the temporal location of the event is *at present*.

So the event that should be considered for temporal annotation is not the one represented by the event anchor itself, but rather the one *described by the whole clause/sentence headed by the event anchor*. This allows all sorts of temporal clues in the same clause/sentence to help decide the temporal location of the event, hence makes the annotation task easier in many cases.

3 Annotation procedure

The annotation process consists of two separate stages, with a different annotation procedure in place for each. The first stage involves only one annotator, and it deals with picking out pairs of event anchors based on the discourse relation as described in Section 2.1. The output of this stage defines the targets for the next stage of annotation: temporal relation annotation. Temporal relation annotation is a two-phase process, including double-blind annotation by two annotators and then adjudication by a judge.

With this procedure in place, the results we report in Section 4 are all from the second stage. Two annotators go through ten weeks of training, which includes annotating 10 files each week, submitting them to adjudication, and then attending a training session at the end of each week. In the training session, the judge discusses with the annotators her adjudication notes from the previous week, as well as specific questions the annotators raise.

The data set consists of 100 files taken from the Chinese Treebank (Xue et al., 2005). The source of these files is Xinhua newswire. The annotation is carried out within the confines of the Brandeis Annotation Tool (BAT)[2] (Verhagen, 2010).

4 Evaluation and discussion

Table 1 reports the inter-annotator agreement of temporal annotation, both between the two annotators (A and B) and between each annotator and the judge (J), over a training period of ten weeks. Each week, 10 files are assigned, averaging about 315 event pairs for annotation.

Table 1 shows that annotators have taken up the temporal annotation scheme fairly quickly, reaching 75% agreement within three weeks. After several

[2]http://timeml.org/site/bat-versions/bat-redesign

Week	No. of tokens	f(A, B)	f(A, J)	f(B, J)
1	310	0.4806		
2	352	0.6278		
3	308	0.7532		
4	243	0.7737		
5	286	0.8007	0.8601	0.8566
6	299	0.7659	0.8662	0.8896
7	296	0.7973	0.8784	0.8784
8	323	0.7988	0.8978	0.8793
9	358	0.8212	0.9106	0.8966
10	378	0.8439	0.9365	0.8995

Table 1: Inter-annotator agreement over 10 weeks of training.

weeks of consolidation and fine-tuning, the agreement slowly reaches the lower 80% towards the end of the 10-week training period. This level of agreement is a substantial improvement over the previously reported results, at 65%, for both English and Chinese data (Verhagen et al., 2009; Xue and Zhou, 2010). This indicates that the general direction of our experiment is on the right track.

Table 2 below is the confusion matrix based on the annotation data from the final 4 weeks:

	a	b	o	na	nb	g	i
a	148	3	19	0	1	0	1
b	0	344	29	1	0	0	7
o	14	10	1354	3	3	2	82
na	0	0	3	3	0	0	0
nb	0	0	1	0	1	0	0
g	2	1	9	0	0	13	1
i	3	7	67	0	0	1	572

Table 2: Confusion matrix on annotation from Weeks 7-10: *a=after; b=before; o=overlap; na=not-after; nb=not-before; g=groupie; i=irrelevant.*

The matrix is fairly clean except when the value "*overlap*" is concerned. This value really stands out in more than one way. It is the most nebulous one in the whole scheme, prone to be confused with all six other values. In particular, it is most likely to be confused with the value "*irrelevant*". It is also the most used value among all seven values, covering roughly half of the tokens. We will discuss this value in more detail in Section 4.2 below.

The value "*groupie*" may also seem troublesome if we look at mis-classification as a percentage of its total occurrences, however, it may not be as bad as it seems. As pointed out in Section 2.2.2, despite the fact that the linguistic phenomenon this value captures can, and does, co-occur with temporal relations represented by other values, we had to set it up as an opposing value to the rest due to technical restrictions. If/when this value is set up as a stand-alone feature to capture the linguistic phenomenon fully, the percentage of mis-classification should drop significantly because the number of total occurrences will increase dramatically.

The overall distribution of values shown in Table 2 is very skewed. At one end of the distribution spectrum is the value *"overlap"*, covering half of the data; at the other end are the values "*not-before*" and "*not-after*", covering less than 0.3% of the token combined. It raises the question if such a classification scheme is well-designed to produce data useful for machine learning.

To shed light on what is behind the numbers and to uncover trends that numbers do not show, we also take a closer look at the annotation data. Three issues stand out.

4.1 Event anchor

In our current scheme, effort is made to pick out the predicate from a clause as the event anchor for temporal annotation. Our experiment suggests maybe this step should be skipped since it, in practice, undermines a specification of the scheme. The specification is that the event to be considered for temporal annotation is the one being described by *the whole clause*, but the practice of displaying a mere word to the annotator in effect instructs the annotator to concentrate on *the word* itself, rather than the clause. Despite repeated reminder during training sessions, the suggestive power of the display still sometimes gets the upper hand. (7) presents such an example concerning *e1* and *e2*.

(7) 在 此 期间 , 西非 维和
PREP this period , West Africa peacekeeping
部队 曾 [$_{e1}$ 出动] 战斗机 轰炸 叛军
force once dispatch fighter jet bomb rebel
阵地 , [$_{e2}$ 炸死] 叛军 约 50余
position , bomb-dead rebel about 50 plus

人。
CL

"During this period, West African Peacekeeping Force [$_{e1}$ dispatched] fighter jets and bombed rebel positions, [$_{e2}$ killing] about 50 rebel troops."

One annotator classified the relation as "*before*", obviously thinking of the event of dispatching fighter jets as *e1*; had he considered the event of dispatching fighter jets *and* bombing the rebel positions, the event being described by the clause, the value would have easily been "*overlap*".

Since displaying the single-word event anchor sometimes leads annotators astray, this step probably should be skipped. Doing so also simplifies the annotation process.

4.2 The value "*overlap*"

As pointed out above, the value "*overlap*" is quite a troubling character in the classification scheme: it is both the most-used and probably the least well-defined. Annotation data show that when it is confused with *"after", "before", "not-after"*, and *"not-before"*, it usually involves a perceptually punctual event ("pp-event" henceforth) and a perceptually lasting event ("pl-event" henceforth), and the issue is whether the pp-event coincides with one of the temporal edges of the pl-event. If it does, then the value is "*overlap*"; otherwise, it is *"after"*/*"before"*. And on top of it is the factor of how sure one is of the issue: if one is sure, either way, the value is "*overlap*"/*"after"*/*"before"*; otherwise, it is *"not-after"*/*"not-before"*. Below is an example on which the two annotators disagree as to whether the relation between *e1* and *e2* should be classified as *"before"* or *"overlap"*.

(8) 此外 , 巴西 女子 国家队 在
in addition , Brazil woman national team PREP
南美 足球赛 上 , [$_{e1}$ 横扫]
S. America soccer match POSTP , sweep
千军 如 卷 席 , [$_{e2}$ 登上] 了
thousand-troop like roll mat , ascend ASP
冠军 宝座 。
champion throne .

"In addition, in the South America Cup, Brazilian Women's national team totally [$_{e1}$ annihilated] all their opponents and [$_{e2}$ ascended] the throne of champion."

In this example, *e2* is the pp-event and *e1* is the pl-event. Depending on when one thinks *e2* happened, either as soon as the last match ended or at the later medal ceremony, (and if the former, whether there is temporal overlap between *e1* and *e2*), it is classified as either "*before*" or "*overlap*; and if one is unsure, it can be classified as "*not-after*".

Such cases again raise the same question as the drastically uneven distribution of values shown in Table 2: *Does the current classification scheme slice the temporal pie the right way?* Let us make a poster child out of *"overlap"*: it seems to both impose too stringent a condition and not make enough distinction. It imposes too stringent a condition on those cases like (8) to which whether there is temporal overlap seems beside the point. At the same time, it does not make enough distinction for cases like (4), in which an event does share one edge of another event temporally: once such cases are classified as *"overlap"*, the specific information regarding the edge is lost. Such information could be very useful in temporal inference. Since it is infeasible to annotate the temporal relation between all events in an article, temporal inference is needed to expand the scope of temporal annotation. For example, if it is known from annotation that *e1* is before *e2* and *e2* is before *e3*, then it can be inferred *e1* is before *e3*. In the case of "*overlap*", whenever it is one of the premises, no inference can be made, but if the "edge" information is supplied, some inferences are possible.

To make finer-grained distinctions in the classification scheme runs counter to the conventional wisdom that a coarser-grained scheme would do a better job handling vagueness. But our experiment has proven the conventional wisdom wrong: our seven-value system achieved much higher agreement than the old six-value system. So the key is not *fewer*, but *better*, distinctions, "better" in the sense that they characterize the data in a more intuitive and insightful way. Temporal relation in natural language is "too" vague only when we judge it against a system of temporal logic, in fact, we think the right word to describe temporal relation in natural language is *"flexible"*: it is as precise as the situation calls for. To characterize the flexibility better, for starters, *"overlap"* needs to be restructured for reasons put forth above, and *"not-before"* and *"not-*

after" should be discarded since they obviously do not carry weight.

4.3 Objective vs. subjective temporal reference

A major contributor to uncertainty and disagreement in annotation is subjective temporal reference. Subjective temporal reference is made based on the author's perspective of the temporal axis, for example, "今天" ("today"), "目前" ("at present), and "过去" ("past"). In this group, references with a fixed span do not constitute a problem once the point of utterance is determined (e.g. literal use of "*today*", "*this month*"); it is those with an elastic temporal span that cause disagreement. For example, *"at present"* can have a span of a second, or several minutes, or a couple of hours, or even years depending on the context. When an event modified with this type of temporal expression is paired with another event modified with direct reference to a point/span on the temporal axis (i.e. with an objective reference), annotation becomes tricky. The event pair *e1-e2* in the two-sentence sequence below is such an example.

(9) 过去 , 在 长江 上 建 大桥
past , PREP Yangtze River POSTP build bridge
是 件 国家大事 , 现今 几乎 [$_{e1}$
be CL national affair , nowadays almost
成为] 平常事。 一九九二年, 江苏
become common scene. 1992-year, Jiangsu
扬中县 农民 [$_{e2}$ 集资]
Yangzhong County farmer raise funds
建成 了 扬中 长江 大桥 , 而
build-finish ASP Yangzhong Yangtze Bridge , and
湖北 的 赤壁 长江 大桥 总 投资
Hubei DE Chibi Yangtze Bridge total invest
三亿多 元 , 全部 靠 民间
300 million plus Yuan , all depend private
集资 建成 。
raise funds build-finish .

"In the past, building a bridge on Yangtze River was a national affair, nowadays it almost [$_{e1}$becomes] a common scene. In 1992, farmers in Yangzhong County, Jiangsu Province [$_{e2}$raised] funds and completed Yangzhong Yangtze Bridge, while Chibi Yangtze Bridge in Hubei Province cost more than 300 million Yuan, all from private fund-raising."

This is taken from a piece written in 1997. In the context, it is clear that the contrast is between the situation before the opening-up of China and the situation about 20 years later. So it is reasonable to assume that the year 1992 falls inside the span of what the author considered *nowadays*; at the same time, it seems also reasonable to assume a narrow interpretation of "现今" ("nowadays") that does not include the year 1992 in the span. These two interpretations would result in "*overlap*" and "*after*" respectively, and actually did so in our experiment.

There are also extreme cases in which objective and subjective temporal references come in direct conflict. For example,

(10) 当 记者 [$_{e1}$ 问及] 中 俄
while reporter ask about China Russia
关系 的 现状 和 合作 前景
relationship DE status and cooperation prospect
时 , 江泽民 主席 [$_{e2}$ 说] , ...
when , Jiang Zemin President say , ...

"When a reporter [$_{e1}$ asked] about the status of China-Russia relationship and the prospects for cooperation, President Jiang Zemin [$_{e2}$ said], ..."

The relation between *e1* and *e2* is *before* based on objective reference, but *overlap* according to the subjective reference, indicated by "当..时" ("when"). This problem should be factored in when a new classification scheme is designed.

5 Conclusions

In this paper, we have described an experiment that focuses on two aspects of the task of annotating temporal relation of main events: annotation target selection and a better-fitting temporal classification scheme. Experiment results show that selecting main event pairs based on discourse structure and modeling the classification scheme after the data improves inter-annotator agreement dramatically. Results also show weakness of the current temporal classification scheme. For that, we propose a restructuring along the lines of what this experiment has proven working: making more intuitive and insightful distinctions that characterize the data better. This direction can be taken to improve other high-level temporal annotation tasks that have been plagued by the same "vagueness" problem.

Acknowledgments

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.

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.

Heng Ji. 2010. Challenges from information extraction to information fusion. In *Proceedings of COLING 2010*, pages 507–515, Beijing, China, August.

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

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber, 2007. *The Penn Discourse Treebank 2.0 Annotation Manual*. The PDTB Research Group, December.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

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.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden, July. Association for Computational Linguistics.

Marc Verhagen. 2010. The Brandeis Annotation Tool. In *Language Resources and Evaluation Conference, LREC 2010*, pages 3638–3643, Malta.

Nianwen Xue and Yuping Zhou. 2010. Applying Syntactic, Semantic and Discourse Constraints to Chinese Temporal Annotation. In *Proceedings of COLING 2010*, pages 1363–1372, Beijing, China, August.

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.

Yuping Zhou and Nianwen Xue. 2011. A PDTB-inspired Discourse Annotation Scheme for Chinese. Submitted to EMNLP 2011.