PDTB-style Discourse Annotation of Chinese Text

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

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

Abstract

We describe a discourse annotation scheme for Chinese and report on the preliminary results. Our scheme, inspired by the Penn Discourse TreeBank (PDTB), adopts the lexically grounded approach; at the same time, it makes adaptations based on the linguistic and statistical characteristics of Chinese text. Annotation results show that these adaptations work well in practice. Our scheme, taken together with other PDTB-style schemes (e.g. for English, Turkish, Hindi, and Czech), affords a broader perspective on how the generalized lexically grounded approach can flesh itself out in the context of cross-linguistic annotation of discourse relations.

1 Introduction

In the realm of discourse annotation, the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008) separates itself by adopting a lexically grounded approach: Discourse relations are lexically anchored by discourse connectives (e.g., *because, but, therefore*), which are viewed as predicates that take abstract objects such as propositions, events and states as their arguments. In the absence of explicit discourse connectives, the PDTB asks the annotator to fill in a discourse connective that best describes the discourse relation between these two sentences, instead of selecting from an inventory of predefined discourse relations. By keeping the discourse annotation lexically grounded even in the case of implicit discourse relations, the PDTB appeals to the annotator's judgment at an intuitive level. This is in contrast with an approach in which the set of discourse relations are pre-determined by linguistic experts and the role of the annotator is just to select from those choices (Mann and Thompson, 1988; Carlson et al., 2003). This lexically grounded approach led to consistent and reliable discourse annotation, a feat that is generally hard to achieve for discourse annotation. The PDTB team reported inter-annotator agreement in the lower 90% for explicit discourse relations (Miltsakaki et al., 2004).

In this paper we describe a discourse annotation scheme for Chinese that adopts this lexically grounded approach while making adaptations when warranted by the linguistic and statistical properties of Chinese text. This scheme is shown to be practical and effective in the annotation experiment.

The rest of the paper is organized as follows: In Section 2, we review the key aspects of the PDTB annotation scheme under discussion in this paper. In Section 3, we first show that some key features of Chinese make adaptations necessary in Section 3.1, and then in Section 3.2, we present our systematic adaptations that follow from the differences outlined in Section 3.1. In Section 4, we present the preliminary annotation results we have so far. And finally in Section 5, we conclude the paper.

2 The PDTB annotation scheme

As mentioned in the introduction, discourse relation is viewed as a predication with two arguments in the framework of the PDTB. To characterize the predication, the PDTB annotates its argument structure and sense. Two types of discourse relation are distinguished in the annotation: explicit and implicit.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 69–77,
Jeju, Republic of Korea, 8-14 July 2012.

Although their annotation is carried out separately, it conforms to the same paradigm of a discourse connective with two arguments. In what follows, we highlight the key points that will be under discussion in the following sections. To get a more comprehensive and detailed picture of the PDTB scheme, see the PDTB 2.0 annotation manual (Prasad et al., 2007).

2.1 Annotation of explicit discourse relations

Explicit discourse relations are those anchored by explicit discourse connectives in text. Explicit connectives are drawn from three grammatical classes:

- **Subordinating conjunctions**: e.g., *because, when, since, although*;
- **Coordinating conjunctions**: e.g., *and, or, nor*;
- **Discourse adverbials**: e.g., *however, otherwise, then, as a result, for example*.

Not all uses of these lexical items are considered to function as a discourse connective. For example, coordinating conjunctions appearing in VP coordinations, such as "*and*" in (1), are not annotated as discourse connectives.

(1) More common chrysotile fibers are curly <u>and</u> are more easily rejected by the body, Dr. Mossman explained.

The text spans of the two arguments of a discourse connective are marked up. The two arguments, *Arg1* and *Arg2*, are defined based on the physical location of the connective: *Arg2* is the argument expressed by the clause syntactically bound to the connective, and *Arg1* is the other argument. There are no restrictions on how many clauses can be included in the text span for an argument other than the *Minimality Principle*: 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.

2.2 Annotation of implicit discourse relations

In the case of implicit discourse relations, annotators are asked to insert a discourse connective that best conveys the implicit relation; when no such connective expression is appropriate, the implicit relation is further distinguished as the following three subtypes:

- **AltLex**: when insertion of a connective leads to redundancy due to the presence of an alternatively lexicalized expression, as in (2).
- **EntRel**: when the only relation between the two arguments is that they describe different aspects of the same entity, as in (3).
- **NoRel**: when neither a lexicalized discourse relation nor entity-based coherence is present. It is to be noted that at least some of the "NoRel" cases are due to the *adjacency constraint* (see below for more detail).

(2) And she further stunned her listeners by revealing her secret garden design method: [$_{Arg1}$ Commissioning a friend to spend five or six thousand dollars . . . on books that I ultimately cut up.] [$_{Arg2}$ <u>*AltLex* After that</u>, the layout had been easy.

(3) [$_{Arg1}$ Hale Milgrim, 41 years old, senior vice president, marketing at Elecktra Entertainment Inc., was named president of Capitol Records Inc., a unit of this entertainment concern]. [$_{Arg2}$ *EntRel* Mr. Milgrim succeeds David Berman, who resigned last month].

There are restrictions on what kinds of implicit relations are subjected to annotation, presented below. These restrictions do not have counterparts in explicit relation annotation.

- Implicit relations between adjacent clauses in the same sentence not separated by a semicolon are not annotated, even though the relation may very well be definable. A case in point is presented in (4) below, involving an intrasentential comma-separated relation between a main clause and a free adjunct.
- Implicit relations between adjacent sentences across a paragraph boundary are not annotated.
- The *adjacency constraint*: At least some part of the spans selected for *Arg1* and *Arg2* must belong to the pair of adjacent sentences initially identified for annotation.

(4) [$_{MC}$ The market for export financing was liberalized in the mid-1980s], [$_{FA}$ forcing the bank to face competition].

2.3 Annotation of senses

Discourse connectives, whether originally present in the data in the case of explicit relations, or filled in by annotators in the case of implicit relations, along with text spans marked as "AltLex", are annotated with respect to their senses. There are three levels in the sense hierarchy:

- **Class**: There are four major semantic classes: TEMPORAL, CONTINGENCY, COMPARISON, and EXPANSION;
- **Type**: A second level of *types* is further defined for each semantic class. For example, under the class CONTINGENCY, there are two types: "Cause" (relating two situations in a direct cause-effect relation) and "Condition" (relating a hypothetical situation with its (possible) consequences);[1]
- **Subtype**: A third level of *subtypes* is defined for some, but not all, types. For instance, under the type "CONTINGENCY:Cause", there are two subtypes: "reason" (for cases like *because* and *since*) and "result" (for cases like *so* and *as a result*).

It is worth noting that a type of implicit relation, namely those labeled as "EntRel", is not part of the sense hierarchy since it has no explicit counterpart.

3 Adapted scheme for Chinese

3.1 Key characteristics of Chinese text

Despite similarities in discourse features between Chinese and English (Xue, 2005), there are differences that have a significant impact on how discourse relations could be best annotated. These differences can be illustrated with (5):

(5) 据悉 ， [$_{AO1}$ 东莞 海关
according to reports , Dongguan Customs
共 接受 企业 合同 备案
in total accept company contract record
八千四百多 份] ， [$_{AO2}$ 比 试点
8400 plus CLASS , compare pilot
前 略 有 上升] ， [$_{AO3}$ 企业
before slight EXIST increase , company
反应 良好] ， [$_{AO4}$ 普遍
respond/response well/good , generally
表示 接受] 。
acknowledge accept/acceptance .

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

This sentence 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 more natural to break them down into two sentences or two semi-colon-separated clauses, but in Chinese, *not only are they merely separated by comma, but also there is no connective relating them.*

This difference in writing style necessitates rethinking of the annotation scheme. If we apply the PDTB scheme to the English translation, regardless of whether the two pieces of facts are expressed in two sentences or two semi-colon-separated clauses, at least one discourse relation will be annotated, relating these two text units. In contrast, if we apply the same scheme to the Chinese sentence, no discourse relation will be picked out because this is just one comma-separated sentence with no explicit discourse connectives in it. In other words, the discourse relation within the Chinese sentence, which would be captured in its English counterpart following the PDTB procedure, would be lost when annotating Chinese. Such loss is not a sporadic occurrence but rather a very prevalent one since it is associated with the customary writing style of Chinese. To ensure a reasonable level of coverage, we need to consider comma-delimited intra-sentential implicit relations when annotating Chinese text.

There are some complications associated with this move. One of them is that it introduces into discourse annotation considerable ambiguity associated with the comma. For example, the first instance of comma in (5), immediately following "据悉" ("according to reports"), clearly does not indicate a discourse relation, so it needs to be spelt out in

[1] There is another dimension to this level, i.e. literal or pragmatic use. If this dimension is taken into account, there could be said to be *four* types: "Cause", "*Pragmatic Cause*", "Condition", and "*Pragmatic Condition*". For details, see Prasad et al. (2007).

the guidelines how to exclude such cases of comma as discourse relation indicators. We think, however, that disambiguating the commas in Chinese text is valuable in its own right and is a necessary step in annotating discourse relations.

Another complication is that some comma-separated chunks are ambiguous as to whether they should be considered potential arguments in a discourse relation. The chunks marked AO2 and AO4 in (5) are examples of such cases. They, judging from their English translation, may seem clear cases of *free adjuncts* in PDTB terms (Prasad et al., 2007), but there is no justification for treating them as such in Chinese. The lack of justification comes from at least three features of Chinese:

- Certain words, for instance, “反应” (“respond/response”), “良好” (“well/good”) and “接受” (“accept/acceptance”), are ambiguous with respect to their POS, and when they combine, the resulting sentence may have more than one syntactic analysis. For example, AO3 may be literally translated as “Companies responded well” or “Companies’ response was good”.
- There are no inflectional clues to differentiate free adjuncts and main clauses. For example, one can be reasonably certain that “表示” (“acknowledge”) functions as a verb in (5), however, there is no indication whether it is in the form corresponding to “acknowledging” or “acknowledged” in English. Or putting it differently, whether one wants to express in Chinese the meaning corresponding to the *-ing* form or the tensed form in English, the same form “表示” could apply.
- Both subject and object can be dropped in Chinese, and they often are when they are inferable from the context. For example, in the two-sentence sequence below, the subject of (7) is dropped since it is clearly the same as the subject of the previous sentence in (6) .

(6) [$_{S1}$ 近 五 年 来 ， 上海
recent five years since , Shanghai through
通过 积极 从 外 省 市
actively from other province city procure
收购 出口 货源 、 举办 中国
export supply , organize China East
华东 出口 商品 交易会 等
Export Commodity Fair etc. event,
活动 ， 增强 口岸 对 全国
strengthen port to whole country DE
的 辐射 能力。]
connection capability .

“[$_{S1}$ In the past five years, Shanghai strengthened the connection of its port to other areas of the country through actively procuring export supplies from other provinces and cities, and through organizing events such as the East China Export Commodities Fair.]”

(7) [$_{S2}$ 同时 ， 发展
At the same time , develop
跨国 经营 ，大力 开拓
transnational operation , vigorously open up
多元化 市场。]
diversified market

“[$_{S2}$ At the same time, (it) developed transnational operations (and) vigorously opened up diversified markets.]”

Since the subject can be omitted from the entire sentence, absence or presence of subject in a clause is not an indication whether the clause is a main clause or a free adjunct, or whether it is part of a VP coordination without a connective. So if we take into account both the lack of differentiating inflectional clues and the possibility of omitting the subject, AO4 in (5) may be literally translated as “generally acknowledging acceptance”, or “(and) generally acknowledged acceptance”, or “(companies) generally acknowledged acceptance”, or “(companies) generally acknowledged (they) accepted (it)”.

Since in Chinese, there is no reliable indicator distinguishing between main clauses and free adjuncts, or distinguishing between coordination on the clause level without the subject and coordination on the VP level, *we will not rely on these distinctions in annotation*, as the PDTB team does in their annotation.

These basic decisions directly based on linguistic characteristics of Chinese lead to more systematic adaptations to the annotation scheme, to which we will turn in the next subsection.

3.2 Systematic adaptations

The main consequence of the basic decisions described in Section 3.1 is that we have a whole lot more tokens of implicit relation than explicit relation to deal with. According to a rough count on 20 randomly selected files from Chinese Treebank (Xue et al., 2005), 82% are tokens of implicit relation, compared to 54.5% in the PDTB 2.0. Given the overwhelming number of implicit relations, we re-examine where it could make an impact in the annotation scheme. There are three such areas.

3.2.1 Procedural division between explicit and implicit discourse relation

In the PDTB, explicit and implicit relations are annotated separately. This is probably partly because explicit connectives are quite abundant in English, and partly because the project evolved in stages, expanding from the more canonical case of explicit relation to implicit relation for greater coverage. When annotating Chinese text, maintaining this procedural division makes much less sense: the landscape of discourse relation (or at least the key elements of it) has already been mapped out by the PDTB work and to set up a separate task to cover 18% of the data does not seem like a worthwhile bother without additional benefits for doing so.

So the question now is *how to annotate explicit and implicit relations in one fell swoop?* In Chinese text, the use of a discourse connective is almost always accompanied by a punctuation or two (usually period and/or comma), preceding or flanking it. So a sensible solution is to rely on punctuations as the denominator between explicit and implicit relations;and in the case of explicit relation, the connective will be marked up as an attribute of the discourse relation. This unified approach simplifies the annotation procedure while preserving the explicit/implicit distinction in the process.

One might question, at this point, whether such an approach can still call itself "lexically grounded". Certainly not if one interprets the term literally ; but in a broader sense, our approach can be seen as an instantiation of a generalized version of it, much the same way that the PDTB is an, albeit different, instantiation of it for English. The thrust of the lexically grounded approach is that discourse annotation should be a data-driven, bottom-up process, rather than a top-down one, trying to fit data into a prescriptive system. Once the insight that a discourse connective functions like a predicate with two arguments is generalized to cover *all* discourse relations, there is no fundamental difference between explicit and implicit discourse relations: both work like a predicate whether or not there is a lexicalization of it. As to what role this distinction plays in the annotation procedure, it is an engineering issue, depending on a slew of factors, among which are cross-linguistic variations. In the case of Chinese, we think it is more economical to treat explicit and implicit relations alike in the annotation process.

To treat explicit and implicit relations alike actually goes beyond annotating them in one pass; it also involves how they are annotated, which we discuss next.

3.2.2 Annotation of implicit discourse relations

In the PDTB, treatment of implicit discourse relations is modeled after that of explicit relations, and at the same time, some restrictions are put on implicit, but not explicit, relations. This is quite understandable: implicit discourse relations tend to be vague and elusive, so making use of explicit relations as a prototype helps pin them down, and restrictions are put in place to strike a balance between high reliability and good coverage. When implicit relations constitute a vast majority of the data as is the case with Chinese, both aspects need to be re-examined to strike a new balance.

In the PDTB, annotators are asked to insert a discourse connective that best conveys the implicit discourse relation between two adjacent discourse units; when no such connective expression is appropriate, the implicit discourse relation is further distinguished as "AltLex", "EntRel", and "NoRel". The inserted connectives and those marked as "AltLex", along with explicit discourse connectives, are further annotated with respect to their senses.

When a connective needs to be inserted in a majority of cases, the difficulty of the task really stands out. In many cases, it seems, there is a good reason for not having a connective present and because of it, the wording rejects insertion of a connective even if it expresses the underlying discourse relation exactly (or sometimes, maybe the wording itself is the reason for not having a connective). So to try to insert a connective expression may very well be too hard a task for annotators, with little to show for their effort in the end.

Furthermore, the inter-annotator agreement for providing an explicit connective in place of an implicit one is computed based on the *type* of explicit connectives (e.g. cause-effect relations, temporal relations, contrastive relations, etc.), rather than based on their identity (Miltsakaki et al., 2004). This suggests that a reasonable degree of agreement for such a task may only be reached with a coarse classification scheme.

Given the above two considerations, our solution is to annotate implicit discourse relations with their senses directly, bypassing the step of inserting a connective expression. It has been pointed out that to train annotators to reason about pre-defined abstract relations with high reliability might be too hard a task (Prasad et al., 2007). This difficulty can be overcome by associating each semantic type with one or two prototypical explicit connectives and asking annotators to consider each to see if it expresses the implicit discourse relation. This way, annotators have a concrete aid to reason about abstract relations without having to choose one connective from a set expressing roughly the same relation or having to worry about whether insertion of the connective is somehow awkward.

It should be noted that annotating implicit relations directly with their senses means that sense annotation is no longer restricted to those that can be lexically expressed, but also includes those that cannot, notably those labeled "EntRel/NoRel" in the PDTB.[2] In other words, we annotate senses of *discourse relations*, not just connectives and their lexical alternatives (in the case of AltLex). This expansion is consistent with the generalized view of the lexically grounded approach discussed in Section 3.2.1.

With respect to restrictions on implicit relation, we will adopt them as they prove to be necessary in the annotation process, with one exception. The exception is the restriction that implicit relations between adjacent clauses in the same sentence not separated by a semi-colon are not annotated. This restriction seems to apply mainly to a main clause and any free adjunct attached to it in English; in Chinese, however, the distinction between a main clause and a free adjunct is not as clear-cut for reasons explained in Section 3.1. So this restriction is not applicable for Chinese annotation.

3.2.3 Definition of Arg1 and Arg2

The third area that an overwhelming number of implicit relation in the data affects is how *Arg1* and *Arg2* are defined. As mentioned in the introduction, discourse relations are viewed as a predication with two arguments. These two arguments are defined based on the physical location of the connective in the PDTB: *Arg2* is the argument expressed by the clause syntactically bound to the connective and *Arg1* is the other argument. In the case of implicit relations, the label is assigned according to the text order.

In an annotation task where implicit relations constitute an overwhelming majority, the distinction of *Arg1* and *Arg2* is meaningless in most cases. In addition, the phenomenon of parallel connectives is predominant in Chinese. Parallel connectives are pairs of connectives that take the same arguments, examples of which in English are "*if..then*", "*either..or*", and "*on the one hand..on the other hand*". In Chinese, most connectives are part of a pair; though some can be dropped from their pair, it is considered "proper" or formal to use both. (8) below presents two such examples, for which parallel connectives are not possible in English.

(8) a. 伦敦 股市 因 适逢
London stock market because coincide
银行节 ，故 没有 开市。
Bank Holiday , therefore NEG open market

"London Stock Market did not open because it was Bank Holiday."

b. 虽然 他们 不 离 土 、不 离
Although they NEG leave land , NEG leave
乡 ，但 严格 来 讲 已
home village , but strict PART speak already
不再 是 传统 意义 上 的 农民。
no longer be tradition sense PREP DE peasant

"Although they do not leave land or their home village, strictly speaking, they are no longer peasants in the traditional sense."

In the PDTB, parallel connectives are annotated discontinuously; but given the prevalence of such phenomenon in Chinese, such practice would generate

[2] Thus "EntRel" and "NoRel" are treated as relation senses, rather than relation types, in our scheme.

a considerably high percentage of essentially repetitive annotation among explicit relations.

So the situation with Chinese is that distinguishing *Arg1* and *Arg2* the PDTB way is meaningless in most cases, and in the remaining cases, it often results in duplication. Rather than abandoning the distinction altogether, we think it makes more sense to define *Arg1* and *Arg2* semantically. It will not create too much additional work beyond distinction of different senses of discourse relation in the PDTB. For example, in the semantic type CONTINGENCY:Cause, we can define "reason" as *Arg1* and "result" as *Arg2*. In this scheme, no matter which one of 因 ("because") and 故 ("therefore") appears without the other, or if they appear as a pair in a sentence, or if the relation is implicit, the *Arg1* and *Arg2* labels will be consistently assigned to the same clauses.

This approach is consistent with the move from annotating senses of *connectives* to annotating senses of *discourse relations*, pointed out in Section 3.2.2. For example, in the PDTB's sense hierarchy, "reason" and "result" are subtypes under type CONTINGENCY:Cause: "reason" applies to connectives like "*because*" and "*since*" while "result" applies to connectives like "*so*" and "*as a result*". When we move to annotating senses of discourse relations, since both types of connectives express the same underlying discourse relation, there will not be further division under CONTINGENCY:Cause, and the "reason"/"result" distinction is an intrinsic property of the semantic type. We think this level of generality makes sense semantically.

4 Annotation experiment

To test our adapted annotation scheme, we have conducted annotation experiments on a modest, yet significant, amount of data and computed agreement statistics.

4.1 Set-up

The agreement statistics come from annotation conducted by two annotators in training so far. The data set consists of 98 files taken from the Chinese Treebank (Xue et al., 2005). The source of these files is Xinhua newswire. The annotation is carried out on the PDTB annotation tool[3].

4.2 Inter-annotator agreement

To evaluate our proposed scheme, we measure agreement on each adaption proposed in Section 3, as well as agreement on argument span determination. Whenever applicable, we also present (roughly) comparable statistics of the PDTB (Miltsakaki et al., 2004). The results are summarized in Table 1.

	Chinese		**PDTB**
	tkn no.	F(p/r) (%)	(%)
rel-ident	3951*	95.4 (96.0/94.7)	N/A
rel-type	3951	95.1	N/A
imp-sns-type	2967	87.4	72
arg-order	3059	99.8	N/A
argument span			
exp-span-xm	1580	84.2	90.2
exp-span-pm	1580	99.6	94.5
imp-span-xm	5934	76.9	85.1
overall-bnd-	14039*	87.7 (87.5/87.9)	N/A

Table 1: Inter-annotator agreement in various aspects of Chinese discourse annotation: *rel-ident*, discourse relation identification; *rel-type*, relation type classification; *imp-sns-type*, classification of sense type of implicit relations; *arg-order*, order determination of Arg1 and Arg2. For agreement on argument spans, the naming convention is <type-of-relation>-<element-as-independent-token>-<matching-method>. *exp*: explicit relations; *imp*: implicit relations; *span*: argument span; *xm*: exact match; *pm*: partial match; *bnd*: boundary. *: number of tokens agreed on by both annotators.

The first adaption we proposed is to annotate explicit and implicit discourse relations in one pass. This introduces two steps, at which agreement can each be measured: First, the annotator needs to make the judgment, at each instance of the punctuations, whether there is a discourse relation (a step we call "relation identification"); second, once a discourse relation is identified, the annotator needs to classify the type as one of "Explicit", "Implicit", or "AltLex" (a step we call "relation type classification"). The agreement at these two steps is 95.4%

[3] http://www.seas.upenn.edu/∼pdtb/tools.shtml#annotator

and 95.1% respectively.

The second adaption is to bypass the step of inserting a connective when annotating an implicit discourse relation and classify the sense directly. The third adaptation is to define Arg1 and Arg2 semantically for each sense. To help annotators think about relation sense abstractly and determine the order of the arguments, we put a helper item alongside each sense label, like "Causation: 因为arg1所以arg2" ("Causation: *because* arg1 *therefore* arg2"). This approach works well, as evidenced by 87.4%[4] and 99.8% agreement for the two processes respectively.

To evaluate agreement on determining argument span, we adopt four measures. In the first three, explicit and implicit relations are calculated separately (although they are actually annotated in the same process) to make our results comparable to the published PDTB results. Each argument span is treated as an independent token and either exact or partial match (i.e. if two spans share one boundary) counts as 1. The fourth measure is less stringent than exact match and more stringent than partial match: It groups explicit and implicit relation together and treats each boundary as an independent token. Typically, an argument span has two boundaries, but it can have four (or more) boundaries when an argument span is interrupted by a connective and/or an AltLex item.

Evidently, determining argument span is the most challenging aspect of discourse annotation. However, it should be pointed out that agreement was on an overall upward trend, which became especially prominent after we instituted a restriction on implicit relations across a paragraph boundary towards the end of the training period. It restricts full annotation to only three specific situations so that most loose and/or hard-to-delimit relations across paragraph boundaries are excluded. This restriction appears to be quite effective, as shown in Table 2.

	num of rel.'s	Overall Arg Span	
		boundary F(p/r) (%)	span-em (%)
last 5 wks	1103	90.0 (90.0/89.9)	80.8
last 3 wks	677	91.0 (91.0/91.0)	82.5
last 2 wks	499	91.8 (91.8/91.8)	84.2

Table 2: Inter-annotator agreement on argument span during the last 5 weeks of training.

5 Conclusions

We have presented a discourse annotation scheme for Chinese that adopts the lexically ground approach of the PDTB while making systematic adaptations motivated by characteristics of Chinese text. These adaptations not only work well in practice, as evidenced by the results from our annotation experiment, but also embody a more generalized view of the lexically ground approach to discourse annotation: Discourse relations are predication involving two arguments; the predicate can be either covert (i.e. Implicit) or overt, lexicalized as discourse connectives (i.e. Explicit) or their more polymorphous counterparts (i.e. AltLex). Consistent with this generalized view is a more semantically motivated sense annotation scheme: Senses of *discourse relations* (as opposed to just connectives) are annotated; and the two arguments of the discourse relation are semantically defined, allowing the sense structure to be more general and less connective-dependent. These framework-level generalizations can be applied to discourse annotation of other languages.

Acknowledgments

This work is supported by the IIS Division of the National Science Foundation via Grant No. 0910532 entitled "Richer Representations for Machine Translation" and by the CNS Division 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

[4] Two more points should be made about this number. First, it may be partially attributed to our differently structured sense hierarchy. It is a flat structure containing the following 12 values: ALTERNATIVE, CAUSATION, CONDITIONAL, CONJUNCTION, CONTRAST, EXPANSION, PROGRESSION, PURPOSE, RESTATEMENT, TEMPORAL, EntRel, and NoRel. Aside from including EntRel and NoRel (the reason and significance of which have been discussed in Section 3.2.2), the revision was by and large not motivated by Chinese-specific features, so we do not address it in detail in this paper. Second, in making the comparison with the PDTB result, the 12-value structure is collapsed into 5 values: TEMPORAL, CONTINGENCY, COMPARISON, EXPANSION, and EntRel/NoRel, which must be different from the 5 values in Miltsakaki et al. (2004), judging from the descriptions.

not necessarily represent the view of the National Science Foundation.

References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Current Directions in Discourse and Dialogue*. Kluwer Academic Publishers.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory. Toward a functional theory of text organization. *Text*, 8(3):243–281.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. Annotating discourse connectives and their arguments. In *Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*, pages 9–16, Boston, MA, May.

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)*.

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.

Nianwen Xue. 2005. Annotating the Discourse Connectives in the Chinese Treebank. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation*, Ann Arbor, Michigan.