Annotating Attribution in the Penn Discourse TreeBank

Rashmi Prasad and **Nikhil Dinesh** and **Alan Lee** and **Aravind Joshi**
University of Pennsylvania
Philadelphia, PA 19104 USA
{rjprasad,nikhild,aleewk,joshi}@linc.cis.upenn.edu

Bonnie Webber
University of Edinburgh
Edinburgh, EH8 9LW Scotland
bonnie@inf.ed.ac.uk

Abstract

An emerging task in text understanding and generation is to categorize information as fact or opinion and to further attribute it to the appropriate source. Corpus annotation schemes aim to encode such distinctions for NLP applications concerned with such tasks, such as information extraction, question answering, summarization, and generation. We describe an annotation scheme for marking the attribution of abstract objects such as propositions, facts and eventualities associated with discourse relations and their arguments annotated in the Penn Discourse TreeBank. The scheme aims to capture the source and degrees of factuality of the abstract objects. Key aspects of the scheme are annotation of the *text spans* signalling the attribution, and annotation of features recording the *source*, *type*, *scopal polarity*, and *determinacy* of attribution.

1 Introduction

News articles typically contain a mixture of information presented from several different perspectives, and often in complex ways. Writers may present information as known to them, or from some other individual's perspective, while further distinguishing between, for example, whether that perspective involves an assertion or a belief. Recent work has shown the importance of recognizing such perspectivization of information for several NLP applications, such as information extraction, summarization, question answering (Wiebe et al., 2004; Stoyanov et al., 2005; Riloff et al., 2005) and generation (Prasad et al., 2005). Part of the goal of such applications is to distinguish between factual and non-factual information, and to identify the source of the information. Annotation schemes (Wiebe et al., 2005; Wilson and Wiebe, 2005; PDTB-Group, 2006) encode such distinctions to facilitate accurate recognition and representation of such perspectivization of information.

This paper describes an extended annotation scheme for marking the attribution of discourse relations and their arguments annotated in the Penn Discourse TreeBank (PDTB) (Miltsakaki et al., 2004; Prasad et al., 2004; Webber et al., 2005), the primary goal being to capture the source and degrees of factuality of abstract objects. The scheme captures four salient properties of attribution: (a) *source*, distinguishing between different types of agents to whom AOs are attributed, (b) *type*, reflecting the degree of factuality of the AO, (c) *scopal polarity* of attribution, indicating polarity reversals of attributed AOs due to surface negated attributions, and (d) *determinacy* of attribution, indicating the presence of contexts canceling the entailment of attribution. The scheme also describes annotation of the *text spans* signaling the attribution. The proposed scheme is an extension of the core scheme used for annotating attribution in the first release of the PDTB (Dinesh et al., 2005; PDTB-Group, 2006). Section 2 gives an overview of the PDTB, Section 3 presents the extended annotation scheme for attribution, and Section 4 presents the summary.

2 The Penn Discourse TreeBank (PDTB)

The PDTB contains annotations of discourse relations and their arguments on the Wall Street Journal corpus (Marcus et al., 1993). Following the approach towards discourse structure in (Webber et al., 2003), the PDTB takes a lexicalized ap-

Proceedings of the Workshop on Sentiment and Subjectivity in Text, pages 31–38,
Sydney, July 2006.

proach towards the annotation of discourse relations, treating *discourse connectives* as the anchors of the relations, and thus as discourse-level predicates taking two *abstract objects* (AOs) as their arguments. For example, in (1), the subordinating conjunction *since* is a discourse connective that anchors a TEMPORAL relation between the event of the earthquake hitting and a state where no music is played by a certain woman. (The 4-digit number in parentheses at the end of examples gives the WSJ file number of the example.)

(1) *She hasn't played any music* since **the earthquake hit**. (0766)

There are primarily two types of connectives in the PDTB: "Explicit" and "Implicit". Explicit connectives are identified form four grammatical classes: subordinating conjunctions (e.g., *because*, *when*, *only because*, *particularly since*), subordinators (e.g., *in order that*), coordinating conjunctions (e.g., *and*, *or*), and discourse adverbials (e.g., *however*, *otherwise*). In the examples in this paper, Explicit connectives are underlined.

For sentences not related by an Explicit connective, annotators attempt to infer a discourse relation between them by *inserting* connectives (called "Implicit" connectives) that *best* convey the inferred relations. For example, in (2), the inferred CAUSAL relation between the two sentences was annotated with *because* as the Implicit connective. Implicit connectives together with their sense classification are shown here in small caps.

(2) *Also unlike Mr. Ruder, Mr. Breeden appears to be in a position to get somewhere with his agenda.* Implicit=BECAUSE (CAUSE) **As a former White House aide who worked closely with Congress, he is savvy in the ways of Washington**. (0955)

Cases where a suitable Implicit connective could not be annotated between adjacent sentences are annotated as either (a) "EntRel", where the second sentence only serves to provide some further description of an entity in the first sentence (Example 3); (b) "NoRel", where no discourse relation or entity-based relation can be inferred; and (c) "AltLex", where the insertion of an Implicit connective leads to *redundancy*, due to the relation being *alternatively lexicalized* by some "non-connective" expression (Example 4).

(3) *C.B. Rogers Jr. was named chief executive officer of this business information concern.* Implicit=EntRel **Mr. Rogers, 60 years old, succeeds J.V. White, 64, who will remain chairman and chairman of the executive committee (0929)**.

(4) *One in 1981 raised to $2,000 a year from $1,500 the amount a person could put, tax-deductible, into the tax-deferred accounts and widened coverage to people under employer retirement plans.* Implicit=AltLex (consequence) **[This caused] an explosion of IRA promotions by brokers, banks, mutual funds and others**. (0933)

Arguments of connectives are simply labelled Arg2, for the argument appearing in the clause syntactically bound to the connective, and Arg1, for the other argument. In the examples here, Arg1 appears in italics, while Arg2 appears in bold.

The basic unit for the realization of an AO argument of a connective is the clause, tensed or untensed, but it can also be associated with multiple clauses, within or across sentences. *Nominalizations* and *discourse deictics* (*this*, *that*), which can also be interpreted as AOs, can serve as the argument of a connective too.

The current version of the PDTB also contains attribution annotations on discourse relations and their arguments. These annotations, however, used the earlier core scheme which is subsumed in the extended scheme described in this paper.

The first release of the Penn Discourse TreeBank, PDTB-1.0 (reported in PDTB-Group (2006)), is freely available from `http://www.seas.upenn.edu/~pdtb`. PDTB-1.0 contains 100 distinct types of Explicit connectives, with a total of 18505 tokens, annotated across the entire WSJ corpus (25 sections). Implicit relations have been annotated in three sections (Sections 08, 09, and 10) for the first release, totalling 2003 tokens (1496 Implicit connectives, 19 AltLex relations, 435 EntRel tokens, and 53 NoRel tokens). The corpus also includes a broadly defined sense classification for the implicit relations, and attribution annotation with the earlier core scheme. Subsequent releases of the PDTB will include Implicit relations annotated across the entire corpus, attribution annotation using the extended scheme proposed here, and fine-grained sense classification for both Explicit and Implicit connectives.

3 Annotation of Attribution

Recent work (Wiebe et al., 2005; Prasad et al., 2005; Riloff et al., 2005; Stoyanov et al., 2005), has shown the importance of recognizing and representing the source and factuality of information in certain NLP applications. Information extraction systems, for example, would perform better

by prioritizing the presentation of factual information, and multi-perspective question answering systems would benefit from presenting information from different perspectives.

Most of the annotation approaches tackling these issues, however, are aimed at performing classifications at either the document level (Pang et al., 2002; Turney, 2002), or the sentence or word level (Wiebe et al., 2004; Yu and Hatzivassiloglou, 2003). In addition, these approaches focus primarily on sentiment classification, and use the same for getting at the classification of facts vs. opinions. In contrast to these approaches, the focus here is on marking attribution on more analytic semantic units, namely the *Abstract Objects* (AOs) associated with predicate-argument discourse relations annotated in the PDTB, with the aim of providing a compositional classification of the factuality of AOs. The scheme isolates four key properties of attribution, to be annotated as features: (1) *source*, which distinguishes between different types of agents (Section 3.1); (2) *type*, which encodes the nature of relationship between agents and AOs, reflecting the degree of factuality of the AO (Section 3.2); (3) *scopal polarity*, which is marked when surface negated attribution reverses the polarity of the attributed AO (Section 3.3), and (4) *determinacy*, which indicates the presence of contexts due to which the entailment of attribution gets cancelled (Section 3.4). In addition, to further facilitate the task of identifying attribution, the scheme also aims to annotate the *text span* complex signaling attribution (Section 3.5)

Results from annotations using the earlier attribution scheme (PDTB-Group, 2006) show that a significant proportion (34%) of the annotated discourse relations have some non-Writer agent as the source for either the relation or one or both arguments. This illustrates the simplest case of the ambiguity inherent for the factuality of AOs, and shows the potential use of the PDTB annotations towards the automatic classification of factuality. The annotations also show that there are a variety of configurations in which the components of the relations are attributed to different sources, suggesting that recognition of attributions may be a complex task for which an annotated corpus may be useful. For example, in some cases, a relation together with its arguments is attributed to the writer or some other agent, whereas in other cases, while the relation is attributed to the writer, one or both of its arguments is attributed to different agent(s). For Explicit connectives. there were 6 unique configurations, for configurations containing more than 50 tokens, and 5 unique configurations for Implicit connectives.

3.1 Source

The *source* feature distinguishes between (a) the writer of the text ("Wr"), (b) some specific agent introduced in the text ("Ot" for other), and (c) some generic source, i.e., some arbitrary ("Arb") individual(s) indicated via a non-specific reference in the text. The latter two capture further differences in the degree of factuality of AOs with non-writer sources. For example, an "Arb" source for some information conveys a higher degree of factuality than an "Ot" source, since it can be taken to be a "generally accepted" view.

Since arguments can get their attribution through the relation between them, they can be annotated with a fourth value "Inh", to indicate that their source value is inherited from the relation.

Given this scheme for *source*, there are broadly two possibilities. In the first case, a relation and both its arguments are attributed to the same source, either the writer, as in (5), or some other agent (here, Bill Biedermann), as in (6). (Attribution feature values assigned to examples are shown below each example; REL stands for the discourse relation denoted by the connective; Attribution text spans are shown boxed.)

(5) Since **the British auto maker became a takeover target last month**, its ADRs have jumped about 78%. (0048)

	REL	Arg1	Arg2
[Source]	Wr	Inh	Inh

(6) "*The public is buying the market* when **in reality there is plenty of grain to be shipped**," said Bill Biedermann . . . (0192)

	REL	Arg1	Arg2
[Source]	Ot	Inh	Inh

As Example (5) shows, text spans for implicit Writer attributions (corresponding to implicit communicative acts such as *I write*, or *I say*), are not marked and are taken to imply Writer attribution by default (see also Section 3.5).

In the second case, one or both arguments have a different source from the relation. In (7), for example, the relation and Arg2 are attributed to the writer, whereas Arg1 is attributed to another agent (here, Mr. Green). On the other hand, in (8) and (9), the relation and Arg1 are attributed to the writer, whereas Arg2 is attributed to another agent.

(7) 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.* (0267)

	REL	Arg1	Arg2
[Source]	Wr	Ot	Inh

(8) *Factory orders and construction outlays were largely flat in December* while purchasing agents said **manufacturing shrank further in October**. (0178)

	REL	Arg1	Arg2
[Source]	Wr	Inh	Ot

(9) *There, on one of his first shopping trips, Mr. Paul picked up several paintings at stunning prices.* . . . Afterward, Mr. Paul is said by Mr. Guterman **to have phoned Mr. Guterman, the New York developer selling the collection, and gloated**. (2113)

	REL	Arg1	Arg2
[Source]	Wr	Inh	Ot

Example (10) shows an example of a generic source indicated by an agentless passivized attribution on Arg2 of the relation. Note that passivized attributions can also be associated with a specific source when the agent is explicit, as shown in (9). "Arb" sources are also identified by the occurrences of adverbs like *reportedly, allegedly*, etc.

(10) Although index arbitrage is said **to add liquidity to markets**, John Bachmann, . . . says *too much liquidity isn't a good thing.* (0742)

	REL	Arg1	Arg2
[Source]	Wr	Ot	Arb

We conclude this section by noting that "Ot" is used to refer to *any* specific individual as the source. That is, no further annotation is provided to indicate *who* the "Ot" agent in the text is. Furthermore, as shown in Examples (11-12), multiple "Ot" sources within the same relation do not indicate whether or not they refer to the same or different agents. However, we assume that the text span annotations for attribution, together with an independent mechanism for named entity recognition and anaphora resolution can be employed to identify and disambiguate the appropriate references.

(11) *Suppression of the book,* Judge Oakes observed, *would operate as a prior restraint and thus involve the First Amendment.* Moreover, and here Judge Oakes went to the heart of the question, **"Responsible biographers and historians constantly use primary sources, letters, diaries, and memoranda**. (0944)

	REL	Arg1	Arg2
[Source]	Wr	Ot	Ot

(12) *The judge was considered imperious, abrasive and ambitious,* those who practiced before him say. Yet, **despite the judge's imperial bearing, no one ever had reason to suspect possible wrongdoing**, says John Bognato, president of Cambria . . . .(0267)

	REL	Arg1	Arg2
[Source]	Wr	Ot	Ot

3.2 Type

The *type* feature signifies the nature of the relation between the agent and the AO, leading to different inferences about the degree of factuality of the AO. In order to capture the factuality of the AOs, we start by making a three-way distinction of AOs into *propositions*, *facts* and *eventualities* (Asher, 1993). This initial distinction allows for a more semantic, compositional approach to the annotation and recognition of factuality. We define the attribution relations for each AO type as follows: (a) *Propositions* involve attribution to an agent of his/her (varying degrees of) commitment towards the truth of a proposition; (b) *Facts* involve attribution to an agent of an evaluation towards or knowledge of a proposition whose truth is taken for granted (i.e., a presupposed proposition); and (c) *Eventualities* involve attribution to an agent of an intention/attitude towards an eventuality. In the case of *propositions*, a further distinction is made to capture the difference in the degree of the agent's commitment towards the truth of the proposition, by distinguishing between "assertions" and "beliefs". Thus, the scheme for the annotation of *type* ultimately uses a four-way distinction for AOs, namely between *assertions*, *beliefs*, *facts*, and *eventualities.* Initial determination of the degree of factuality involves determination of the type of the AO.

AO types can be identified by well-defined semantic classes of verbs/phrases anchoring the attribution. We consider each of these in turn.

Assertions are identified by "assertive predicates" or "verbs of communication" (Levin, 1993) such as *say*, *mention*, *claim*, *argue*, *explain* etc. They take the value "Comm" (for verbs of Communication). In Example (13), the Ot attribution on Arg1 takes the value "Comm" for *type*. Implicit writer attributions, as in the relation of (13), also take (the default) "Comm". Note that when an argument's attribution source is not inherited (as in Arg1 in this example) it also takes its own independent value for *type*. This example thus conveys that there are two different attributions expressed within the discourse relation, one for the relation and the other for one of its arguments, and that both involve assertion of propositions.

(13) 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.* (0267)

	REL	Arg1	Arg2
[Source]	Wr	Ot	Inh
[Type]	Comm	Comm	Null

In the absence of an independent occurrence of attribution on an argument, as in Arg2 of Example (13), the "Null" value is used for the *type* on the argument, meaning that it needs to be derived by independent (here, undefined) considerations under the scope of the relation. Note that unlike the "Inh" value of the *source* feature, "Null" does not indicate inheritance. In a subordinate clause, for example, while the relation denoted by the subordinating conjunction may be asserted, the clause content itself may be presupposed, as seems to be the case for the relation and Arg2 of (13). However, we found these differences difficult to determine at times, and consequently leave this undefined in the current scheme.

Beliefs are identified by "propositional attitude verbs" (Hintikka, 1971) such as *believe*, *think*, *expect*, *suppose*, *imagine*, etc. They take the value "PAtt" (for Propostional Attitude). An example of a belief attribution is given in (14).

(14) Mr. Marcus believes *spot steel prices will continue to fall through early 1990* **and** then **reverse themselves**. (0336)

	REL	Arg1	Arg2
[Source]	Ot	Inh	Inh
[Type]	PAtt	Null	Null

Facts are identified by the class of "factive and semi-factive verbs" (Kiparsky and Kiparsky, 1971; Karttunen, 1971) such as *regret*, *forget*, *remember*, *know*, *see*, *hear* etc. They take the value "Ftv" (for Factive) for *type* (Example 15). In the current scheme, this class does not distinguish between the true factives and semi-factives, the former involving an attitute/evaluation towards a fact, and the latter involving knowledge of a fact.

(15) The other side, he argues knows *Giuliani has always been pro-choice,* even though **he has personal reservations**. (0041)

	REL	Arg1	Arg2
[Source]	Ot	Inh	Inh
[Type]	Ftv	Null	Null

Lastly, *eventualities* are identified by a class of verbs which denote three kinds of relations between agents and eventualities (Sag and Pollard, 1991). The first kind is anchored by *verbs of influence* like *persuade*, *permit*, *order*, and involve one agent influencing another agent to perform (or not perform) an action. The second kind is anchored by *verbs of commitment* like *promise*, *agree*, *try*, *intend*, *refuse*, *decline*, and involve an agent committing to perform (or not perform) an action. Finally, the third kind is anchored by *verbs of orientation* like *want*, *expect*, *wish*, *yearn*, and involve desire, expectation, or some similar mental orientation towards some state(s) of affairs. These sub-distinctions are not encoded in the annotation, but we have used the definitions as a guide for identifying these predicates. All these three types are collectively referred to and annotated as *verbs of control*. *Type* for these classes takes the value "Ctrl" (for Control). Note that the syntactic term *control* is used because these verbs denote uniform structural control properties, but the primary basis for their definition is nevertheless semantic. An example of the control attribution relation anchored by a verb of influence is given in (16).

(16) Eward and Whittington had planned to leave the bank earlier, but Mr. Craven had persuaded them *to remain* until **the bank was in a healthy position**. (1949)

	REL	Arg1	Arg2
[Source]	Ot	Inh	Inh
[Type]	Ctrl	Null	Null

Note that while our use of the term *source* applies literally to agents responsible for the truth of a proposition, we continue to use the same term for the agents for facts and eventualities. Thus, for facts, the *source* represents the bearers of attitudes/knowledge, and for considered eventualities, the *source* represents intentions/attitudes.

3.3 Scopal Polarity

The *scopal polarity* feature is annotated on relations and their arguments to primarily identify cases when verbs of attribution are negated on the surface - syntactically (e.g., *didn't say*, *don't think*) or lexically (e.g., *denied*), but when the negation in fact reverses the polarity of the attributed relation or argument content (Horn, 1978). Example (17) illustrates such a case. The 'but' clause entails an interpretation such as "I think it's not a main consideration", for which the negation must take narrow scope over the embedded clause rather than the higher clause. In particular, the interpretation of the CONTRAST relation denoted by *but* requires that Arg2 should be interpreted under the scope of negation.

(17) *"Having the dividend increases is a supportive element in the market outlook,* but I don't think **it's a main consideration**," he says. (0090)

	REL	Arg1	Arg2
[Source]	Ot	Inh	Inh
[Type]	Comm	Null	PAtt
[Polarity]	Null	Null	Neg

To capture such entailments with surface negations on attribution verbs, an argument of a connective is marked "Neg" for *scopal polarity* when the interpretation of the connective requires the surface negation to take semantic scope over the lower argument. Thus, in Example (17), *scopal polarity* is marked as "Neg" for Arg2.

When the neg-lowered interpretations are not present, *scopal polarity* is marked as the default "Null" (such as for the relation and Arg1 of Example 17).

It is also possible for the surface negation of attribution to be interpreted as taking scope over the relation, rather than an argument. We have not observed this in the corpus yet, so we describe this case with the constructed example in (18). What the example shows is that in addition to entailing (18b) - in which case it would be annotated parallel to Example (17) above - (18a) can also entail (18c), such that the negation is intrepreted as taking semantic scope over the "relation" (Lasnik, 1975), rather than one of the arguments. As the *scopal polarity* annotations for (18c) show, lowering of the surface negation to the relation is marked as "Neg" for the *scopal polarity* of the relation.

(18) a. John doesn't think *Mary will get cured* because **she took the medication**.

b. $\models$ John thinks *that* because **Mary took the medication**, *she will not get cured.*

	REL	Arg1	Arg2
[Source]	Ot	Inh	Inh
[Type]	PAtt	Null	Null
[Polarity]	Null	Neg	Null

c. $\models$ John thinks *that Mary will get cured* not because **she took the medication** (but because she has started practising yoga.)

	REL	Arg1	Arg2
[Source]	Ot	Inh	Inh
[Type]	PAtt	Null	Null
[Polarity]	Neg	Null	Null

We note that *scopal polarity* does not capture the appearance of (opaque) internal negation that may appear on arguments or relations themselves. For example, a modified connective such as *not because* does not take "Neg" as the value for *scopal polarity*, but rather "Null". This is consistent with our goal of marking *scopal polarity* only for lowered negation, i.e., when surface negation from the attribution is lowered to either the relation or argument for interpretation.

3.4 Determinacy

The *determinacy* feature captures the fact that the entailment of the attribution relation can be made indeterminate in context, for example when it appears syntactically embedded in negated or conditional contexts.. The annotation attempts to capture such indeterminacy with the value "Indet". Determinate contexts are simply marked as the default "Null". For example, the annotation in (19) conveys the idea that the belief or opinion about the effect of higher salaries on teachers' performance is not really attributed to anyone, but is rather only being conjectured as a possibility.

(19) It is silly libel on our teachers to think *they would educate our children better* if only **they got a few thousand dollars a year more**. (1286)

	REL	Arg1	Arg2
[Source]	Ot	Inh	Inh
[Type]	PAtt	Null	Null
[Polarity]	Null	Null	Null
[Determinacy]	Indet	Null	Null

3.5 Attribution Spans

In addition to annotating the properties of attribution in terms of the features discussed above, we also propose to annotate the *text span* associated with the attribution. The text span is annotated as a single (possibly discontinuous) complex reflecting three of the annotated features, namely *source*, *type* and *scopal polarity*. The attribution span also includes all non-clausal modifiers of the elements contained in the span, for example, adverbs and appositive NPs. Connectives, however, are excluded from the span, even though they function as modifiers. Example (20) shows a discontinuous annotation of the attribution, where the parenthetical *he argues* is excluded from the attribution phrase *the other side knows*, corresponding to the factive attribution.

(20) The other side, he argues knows *Giuliani has always been pro-choice,* even though **he has personal reservations**. (0041)

	REL	Arg1	Arg2
[Source]	Ot	Inh	Inh
[Type]	Ftv	Null	Null
[Polarity]	Null	Null	Null
[Determinacy]	Null	Null	Null

Inclusion of the fourth feature, *determinacy*, is not "required" to be included in the current scheme because the entailment cancelling contexts

can be very complex. For example, in Example (19), the conditional interpretation leading to the indeterminacy of the relation and its arguments is due to the syntactic construction type of the entire sentence. It is not clear how to annotate the indeterminacy induced by such contexts. In the example, therefore, the attribution span only includes the anchor for the *type* of the attribution.

Spans for implicit writer attributions are left unmarked since there is no corresponding text that can be selected. The absence of a span annotation is simply taken to reflect writer attribution, together with the "Wr" value on the source feature.

Recognizing attributions is not trivial since they are often left unexpressed in the sentence in which the AO is realized, and have to be inferred from the prior discourse. For example, in (21), the relation together with its arguments in the third sentence are attributed to Larry Shapiro, but this attribution is implicit and must be inferred from the first sentence.

(21) "There are certain cult wines that can command these higher prices," says Larry Shapiro of Marty's, . . . "What's different is that it is happening with young wines just coming out. *We're seeing it* partly because **older vintages are growing more scarce**." (0071)

	REL	Arg1	Arg2
[Source]	Ot	Inh	Inh

The spans for such implicit "Ot" attributions mark the text that provides the inference of the implicit attribution, which is just the closest occurrence of the explicit attribution phrase in the prior text.

The final aspect of the span annotation is that we also annotate non-clausal phrases as the anchors attribution, such as prepositional phrases like *according to X*, and adverbs like *reportedly*, *allegedly*, *supposedly*. One such example is shown in (22).

(22) *No foreign companies bid on the Hiroshima project, according to the bureau*. But **the Japanese practice of deep discounting often is cited by Americans as a classic barrier to entry in Japan's market**. (0501)

	REL	Arg1	Arg2
[Source]	Wr	Ot	Inh
[Type]	Comm	Comm	Null
[Polarity]	Null	Null	Null
[Determinacy]	Null	Null	Null

Note that adverbials are free to pick their own *type* of attribution. For example, *supposedly* as an attribution adverb picks "PAtt" as the value for *type*.

3.6 Attribution of Implicit Relations

Implicit connectives and their arguments in the PDTB are also marked for attribution. Implicit connectives express relations that are inferred by the reader. In such cases, the writer intends for the reader to infer a discourse relation. As with Explicit connectives, implicit relations intended by the writer of the article are distinguished from those intended by some other agent introduced by the writer. For example, while the implicit relation in Example (23) is attributed to the writer, in Example (24), both Arg1 and Arg2 have been expressed by someone else whose speech is being quoted: in this case, the implicit relation is attributed to the other agent.

(23) *The gruff financier recently started socializing in upper-class circles.* Implicit = FOR EXAMPLE (ADD.INFO) Although he says he wasn't keen on going, **last year he attended a New York gala where his daughter made her debut**. (0800)

	REL	Arg1	Arg2
[Source]	Wr	Inh	Inh
[Type]	Comm	Null	Null
[Polarity]	Null	Null	Null
[Determinacy]	Null	Null	Null

(24) "*We asked police to investigate why they are allowed to distribute the flag in this way.* Implicit=BECAUSE (CAUSE) **It should be considered against the law**," said Danny Leish, a spokesman for the association.

	REL	Arg1	Arg2
[Source]	Ot	Inh	Inh
[Type]	Comm	Null	Null
[Polarity]	Null	Null	Null
[Determinacy]	Null	Null	Null

For implicit relations, attribution is also annotated for AltLex relations but not for EntRel and NoRel, since the former but not the latter refer to the presense of discourse relations.

4 Summary

In this paper, we have proposed and described an annotation scheme for marking the attribution of both explicit and implicit discourse connectives and their arguments in the Penn Discourse TreeBank. We discussed the role of the annotations for the recognition of factuality in natural language applications, and defined the notion of attribution. The scheme was presented in detail with examples, outlining the "feature-based annotation" in terms of the *source*, *type*, *scopal polarity*, and *determinacy* associated with attribution, and the "span annotation" to highlight the text reflecting the attribution features.

Acknowledgements

The Penn Discourse TreeBank project is partially supported by NSF Grant: Research Resources, EIA 02-24417 to the University of Pennsylvania (PI: A. Joshi). We are grateful to Lukasz Abramowicz and the anonymous reviewers for useful comments.

References

Nicholas. Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht.

Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Attribution and the (non)-alignment of syntactic and discourse arguments of connectives. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, Ann Arbor, Michigan.

Jaakko Hintikka. 1971. Semantics for propositional attitudes. In L. Linsky, editor, *Reference and Modality*, pages 145–167. Oxford.

Laurence Horn. 1978. Remarks on neg-raising. In Peter Cole, editor, *Syntax and Semantics 9: Pragmatics*. Academic Press, New York.

Lauri Karttunen. 1971. Some observations on factivity. *Papers in Linguistics*, 4:55–69.

Carol Kiparsky and Paul Kiparsky. 1971. Fact. In D. D. Steinberg and L. A. Jakobovits, editors, *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology*, pages 345–369. Cambridge University Press, Cambridge.

Howard Lasnik. 1975. On the semantics of negation. In *Contemporary Research in Philosophical Logic and Linguistic Semantics*, pages 279–313. Dordrecht: D. Reidel.

Beth Levin. 1993. *English Verb Classes And Alternations: A Preliminary Investigation*. University of Chicago Press.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

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.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pages 79–86.

Rashmi Prasad, Eleni Miltsakaki, Aravind Joshi, and Bonnie Webber. 2004. Annotation and data mining of the Penn Discourse Treebank. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 88–95, Barcelona, Spain.

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. 2005. The Penn Discourse TreeBank as a resource for natural language generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for NLG*.

Ellen Riloff, Janyce Wiebe, and Willian Phillips. 2005. Exploiting subjectivity classification to improve information extraction. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-2005)*.

Ivan A. Sag and Carl Pollard. 1991. An integrated theory of complement control. *Language*, 67(1):63–113.

The PDTB-Group. 2006. The Penn Discourse TreeBank 1.0 Annotation Manual. Technical Report IRCS-06-01, Institute for Research in Cognitive Science, University of Pennsylvania.

Veseli Stoyanov, Claire Cardie, and Janyce Wiebe. 2005. Multi-perspective question answering using the OpQA corpus. In *Proceedings of HLT-EMNLP*.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL 2002*, pages 417–424.

Bonnie Webber, Aravind Joshi, M. Stone, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.

Bonnie Webber, Aravind Joshi, Eleni Miltsakaki, Rashmi Prasad, Nikhil Dinesh, Alan Lee, and K. Forbes. 2005. A short introduction to the PDTB. In *Copenhagen Working Papers in Language and Speech Processing*.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Janyce Wiebe, Theresa Wilson, , and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2).

Theresa Wilson and Janyce Wiebe. 2005. Annotating attributions and private states. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, Ann Arbor, Michigan.

Hon Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP-2003*, pages 129–136, Saporo, Japan.