Panel Session: Discourse Annotation

Manfred Stede
Dept. of Linguistics
University of Potsdam
stede@ling.uni-potsdam.de

Janyce Wiebe
Dept. of Computer Science
University of Pittsburgh
wiebe@cs.pitt.edu

Eva Hajičová
Faculty of Math. and Physics
Charles University
hajicova@ufal.ms.mff.cuni.cz

Brian Reese
Dept. of Linguistics
Univ. of Texas at Austin
bjreese@mail.utexas.edu

Simone Teufel
Computer Laboratory
Univ. of Cambridge
sht25@cl.cam.uk

Bonnie Webber
School of Informatics
Univ. of Edinburgh
bonnie@inf.ed.ac.uk

Theresa Wilson
Dept. of Comp. Science
Univ. of Pittsburgh
twilson@cs.pitt.edu

1 Introduction

The classical "success story" of corpus annotation are the various syntax treebanks that provide structural analyses of sentences and have enabled researchers to develop a range of new and highly successful data-oriented approaches to sentence parsing. In recent years, however, a number of corpora have been constructed that provide annotations on the *discourse* level, i.e. information that reaches beyond the sentence boundaries. Phenomena that have been annotated include coreference links, the scope of connectives, and coherence relations. Many of these are phenomena on whose handling there is not a general agreement in the research community, and therefore the question of "recycling" corpora by other people and for other purposes is often difficult. (To some extent, this is due to the fact that discourse annotation deals "only" with surface reflections of underlying, abstract objects.) At the same time, the efforts needed for building high-quality discourse corpora are considerable, and thus one should be careful in deciding how to invest those efforts. One aspect of providing added-value with annotation projects is that of *shared* corpora: If a variety of annotation efforts is executed on the same primary data, the series of annotation levels can yield insights that the creators of the individual levels had not explicitly planned for. A clear case is the relationship between coherence relations and connective use: When both levels are marked individually and with independent annotation guidelines, then afterwards the correlations between coherence relations, cue usage (and possibly other factors, if annotated) can be studied systematically. This conception of *multi-level* annotation presupposes, of course, that the technical problems of setting annotation levels in correspondence to one another be resolved.

The panel on discourse annotation is organized by Manfred Stede and Janyce Wiebe. It aims at surveying the scene of discourse corpora, exploring chances for synergy, and identifying desiderata for future corpus creation projects. In preparation for the panel, the participants have provided the following short descriptions of the various copora in whose construction they have been involved.

2 Prague Dependency Treebank (Eva Hajičová, Prague)

One of the maxims of the work on the Prague Dependency Treebank is that one should not overlook, disregard and thus lose what the *sentence* structure offers when one attempts to analyze the structure of discourse, thus moving from "the trees" to "the forest". Therefore, we emphasize that discourse annotation should make use of every possible detail the annotation of the component parts of the discourse, namely the sentences, puts at our disposal. This is, of course, not only true for the surface shape of the sentence (i.e., the surface means of expression), but (and most importantly) for the underlying representation of sentences. The panel contribution will introduce the (multilayered) annotation scenario of the Prague Dependency Treebank and illustrate the point using some of the particular features of the underlying structure of sentences that can be made use of in planning the scenario of discourse 'treebanks'.

Proceedings of the Linguistic Annotation Workshop, pages 191–196,
Prague, June 2007.

3 SDRT in Newspaper Text (Brian Reese, Austin)

We are currently working under the auspices of an NSF grant to build and train a discourse parser and codependent anaphora resolution program to test discourse theories empirically. The training requires the construction of a corpus annotated with discourse structure and coreference information. So far, we have annotated the MUC6[1] corpus for discourse structure and are in the process of annotating the ACE2[2] corpus; both corpora are already annotated for coreference. One of the goals of the project is to investigate whether using the right frontier constraint improves the system's performance in resolving anaphors. Here we detail some experiences we have had with the discourse annotation process.

An implementation of the extant SDRT (Asher and Lascarides, 2003) glue logic for building discourse structures is insufficient to deal with open domain text, and we cannot envision an extended version at the present time able to deal with the problem. Thus, we have opted for a machine learning based approach to discourse parsing based on superficial features, like BNL. To build an implementation to test these ideas, we have had to devise a corpus of texts annotated for discourse structure in SDRT.

Each of the 60 texts in the MUC6 corpus, and now 18 of the news stories in ACE2, were annotated by two people familiar with SDRT. The annotators then conferred and agreed upon a gold standard. Our annotation effort took the hierarchical structure of SDRT seriously and built graphs in which the nodes are discourse units and the arcs represent discourse relations between the units. The units could either be simple (elementary discourse units: EDUs) or they could be complex. We assumed that in principle the units were recursively generated and could have an arbitrary though finite degree of complexity.

4 Potsdam Commentary Corpus (Manfred Stede, Potsdam)

Construction of the Potsdam Commentary Corpus (PCC) began in 2003 and is still ongoing. It is a genre-specific corpus of German newspaper commentaries, taken from the daily papers *Märkische Allgemeine Zeitung* and *Tagesspiegel*. One central aim is to provide a tool for studying mechanisms of argumentation and how they are reflected on the linguistic surface. The corpus on the one hand is a collection of "raw" data, which is used for genre-oriented statistical explorations. On the other hand, we have identified two sub-corpora that are subject to a rich multi-level annotation (MLA).

The *PCC176* (Stede, 2004) is a sub-corpus that is available upon request for research purposes. It consists of 176 relatively short commentaries (12-15 sentences), with 33.000 tokens in total. The sentences have been PoS-tagged automatically (and manually checked); sentence syntax was annotated semi-automatically using the TIGER scheme (Brants et al., 2002) and Annotate[3] tool. In addition, we annotated coreference (PoCos (Krasavina and Chiarcos, 2007)) and rhetorical structure according to RST (Mann and Thompson, 1988). Our annotation software architecture consists of a variety of standard, external tools that can be used effectively for the different annotation types. Their XML output is then automatically converted to a generic format (PAULA, (Dipper, 2005)), which is read into the linguistic database ANNIS (Dipper et al., 2004), where the annotations are aligned, so that the data can be viewed and queried across annotation levels.

The *PCC10* is a sub-corpus of 10 commentaries that serves as "testbed" for further developing the annotation levels. On the one hand, we are applying recent guidelines on annotation of information structure (Götze et al., 2007). On the other hand, based on experiences with the RST annotation, we are replacing the rhetorical trees with a set of distinct, simpler annotation layers: thematic structure, conjunctive relations (Martin, 1992), and argumentation structure (Freeman, 1991); these are complemented by the other levels mentioned above for the PCC176. The primary motivation for this step is the high degree of arbitrariness that annotators reported when producing the RST trees (see (Stede, 2007)). By separating the thematic from the intentional information, and accounting for the surface-oriented

[1]The Message Understanding Conference, www-nlpir.nist.gov/related_projects/muc/.

[2]The Automated Content Extraction program, www.nist.gov/speech/tests/ace/.

[3]www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html

conjunctive relations (which are similar to what is annotated in the PDTB, see Section 6), we hope to

- make annotation easier: handling several "simple" levels individually should be more effective than a single, very complex annotation step;
- end up with less ambiguity in the annotations, since the reasons for specific decisions can be made explicit (by annotations on "simpler" levels);
- be more explicit than a single tree can be: if a discourse fulfills, for example, a function both for thematic development and for the writer's intention, they can both be accounted for;
- provide the central information that a "traditional" rhetorical tree conveys, without loosing essential information.

5 AZ Corpus (Simone Teufel, Cambridge)

The Argumentative Zoning (AZ) annotation scheme (Teufel, 2000; Teufel and Moens, 2002) is concerned with marking argumentation steps in scientific articles. One example for an argumentation step is the description of the research goal, another an overt comparison of the authors' work with rival approaches. In our scheme, these argumentation steps have to be associated with text spans (sentences or sequences of sentences). AZ–Annotation is the labelling of each sentence in the text with one of these labels (7 in the original scheme in (Teufel, 2000)). The AZ labels are seen as relations holding between the meanings of these spans, and the rhetorical act of the entire paper. (Teufel et al., 1999) reports on interannotator agreement studies with this scheme.

There is a strong interrelationship between the argumentation in a paper, and the citations writers use to support their argument. Therefore, a part of the computational linguistics corpus has a second layer of annotation, called CFC (Teufel et al., 2006) or Citation Function Classification. CFC– annotation records for each citation which rhetorical function it plays in the argument. This is following the spirit of research in citation content analysis (e.g., (Moravcsik and Murugesan, 1975)). An example for a citation function would be "motivate that the method used is sound". The annotation scheme contains 12 functions, clustered into "superiority", "neutral comparison/contrast", "praise or usage" and "neutral".

One type of research we hope to do in the future is to study the relationship between these rhetorical phonemena with more traditional discourse phenomena, e.g. anaphoric expressions.

The CmpLg/ACL Anthology corpora consist of 320/9000 papers in computational linguistics. They are partially annotated with AZ and CFC markup. A subcorpus of 80 parallelly annotated papers (AZ and CFF) can be obtained from us for research (12000 sentences, 1756 citations). We are currently porting both schemes to chemistry in the framework of the EPSRC-sponsored project SciBorg. In the course of this work a larger, more general AZ annotation scheme was developed. The SciBorg effort will result in an AZ/CFC–annotated chemistry corpus available to the community in 2009.

In terms of challenges, the most time-consuming aspects of creating this annotated corpus were format conversions on the corpora, and cyclic adaptations of scheme and guidelines. Another problem is the simplification of annotating only full sentences; sometimes, annotators would rather mark a clause or sometimes even just an NP. However, we found these cases to be relatively rare.

6 Penn Discourse Treebank (Bonnie Webber, Edinburgh)

The Penn Discourse TreeBank (Miltsakaki et al., 2004; Prasad et al., 2004; Webber, 2005) annotates *discourse relations* over the Wall Street Journal corpus (Marcus et al., 1993), in terms of *discourse connectives* and their arguments. Following the approach towards discourse structure in (Webber et al., 2003), the PDTB takes a lexicalized approach, treating discourse connectives as the anchors of the relations and thus as discourse-level predicates taking two *Abstract Objects* as their arguments. Annotated are the *text spans* that give rise to these arguments. There are primarily two types of connectives in the PDTB: *explicit* and *implicit*, the latter being *inserted* between adjacent paragraph-internal sentence pairs not related by an explicit connective.

Also annotated in the PDTB is the *attribution* of each discourse relation and of its arguments (Dinesh et al., 2005; Prasad et al., 2007). (Attribution itself is not considered a discourse relation.) A preliminary version of the PDTB was released in April 2006 (PDTB-Group, 2006), and is available for download at http://www.seas.upenn.edu/˜pdtb. This release only has implicit connectives annotated in three sections of the corpus. The annotation of all implicit connectives, along with a hierarchical semantic classification of all connectives (Miltsakaki et al., 2005), will appear in the final release of the PDTB in August 2007.

Here I want to mention three of the challenges we have faced in developing the PDTB:

(I) Words and phrases that can function as connectives can also serve other roles. (Eg, *when* can be a relative pronoun, as well as a subordinating conjunction.) It has been difficult to identify all and only those cases where a token functions as a discourse connective, and in many cases, the syntactic analysis in the Penn TreeBank (Marcus et al., 1993) provides no help. For example, is *as though* always a subordinating conjunction (and hence a connective) or do some tokens simply head a manner adverbial (eg, *seems as though ...* versus *seems more rushed as though ...*)? Is *also* sometimes a discourse connective relating two abstract objects and other times, an adverb that presupposes that a particular property holds of some other entity? If so, when one and when the other? In the PDTB, annotation has erred on the side of false positives.

(II) In annotating implicit connectives, we discovered systematic non-lexical indicators of discourse relations. In English, these include cases of marked syntax (eg, *Had I known the Queen would be here, I would have dressed better.*) and cases of sentence-initial PPs and adjuncts with anaphoric or deictic NPs such as *at the other end of the spectrum*, *adding to that speculation*. These cases labelled ALTLEX, for "alternative lexicalisation" have not been annotated as connectives in the PDTB because they are fully productive (ie, not members of a more easily annotated closed set of tokens). They comprise about 1% of the cases the annotators have considered. Future discourse annotation will benefit from further specifying the types of these cases.

(III) The way in which spans are annotated as arguments to connectives also raises a challenge. First, because the PDTB annotates both structural and anaphoric connectives (Webber et al., 2003), a span can serve as argument to >1 connective. Secondly, unlike in the RST corpus (Carlson et al., 2003) or the Discourse GraphBank (Wolf and Gibson, 2005), discourse segments are not separately annotated, with annotators then identifying what discourse relations hold between them. Instead, in annotating arguments, PDTB annotators have selected the *minimal* clausal text span needed to interpret the relation. This could comprise an embedded, subordinate or coordinate clause, an entire sentence, or a (possibly disjoint) sequence of sentences. As a result, there are fairly complex patterns of spans within and across sentences that serve as arguments to different connectives, and there are parts of sentences that don't appear within the span of *any* connective, explicit or implicit. The result is that the PDTB provides only a partial but complexly-patterned cover of the corpus. Understanding what's going on and what it implies for discourse structure (and possibly syntactic structure as well) is a challenge we're currently trying to address (Lee et al., 2006).

7 MPQA Opinion Corpus (Theresa Wilson, Pittsburgh)

Our opinion annotation scheme (Wiebe et al., 2005) is centered on the notion of *private state*, a general term that covers opinions, beliefs, thoughts, sentiments, emotions, intentions and evaluations. As Quirk et al. (1985) define it, a *private state* is a state that is not open to objective observation or verification. We can further view private states in terms of their functional components — as states of *experiencers* holding *attitudes*, optionally toward *targets*. For example, for the private state expressed in the sentence *John hates Mary*, the experiencer is *John*, the attitude is *hate*, and the target is *Mary*.

We create private state frames for three main types of private state expressions (*subjective expressions*) in text:

- explicit mentions of private states, such as "fears" in "The U.S. fears a spill-over"

- speech events expressing private states, such as "said" in "The report is **full of absurdities**," Xirao-Nima said.

- expressive subjective elements, such as "full of absurdities" in the sentence just above.

Frames include the source (experiencer) of the private state, the target, and various properties such as polarity (*positive*, *negative*, or *neutral*) and intensity (*high*, *medium*, or *low*). Sources are *nested*. For example, for the sentence "China criticized the U.S. report's criticism of China's human rights record", the source is ⟨*writer, China, U.S. report*⟩, reflecting the facts that the writer wrote the sentence and the U.S. report's criticism is the target of China's criticism. It is common for multiple frames to be created for a single clause, reflecting various levels of nesting and the type of subjective expression.

The annotation scheme has been applied to a corpus, called the "Multi-Perspective Question Answering (MPQA) Corpus," reflecting its origins in the 2002 NRRC Workshop on Multi-Perspective Question Answering (MPQA) (Wiebe et al., 2003) sponsored by ARDA AQUAINT (it is also called "OpinionBank"). It contains 535 documents and a total of 11,114 sentences. The articles in the corpus are from 187 different foreign and U.S. news sources, dating from June 2001 to May 2002. Please see (Wiebe et al., 2005) and Theresa Wilson's forthcoming PhD dissertation for further information, including the results of inter-coder agreement studies.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In J. van Kuppevelt & R. Smith, editor, *Current Directions in Discourse and Dialogue*. Kluwer, New York.

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 *ACL Workshop on Frontiers in Corpus Annotation*, Ann Arbor MI.

Stefanie Dipper, Michael Götze, Manfred Stede, and Tillmann Wegst. 2004. Annis: A linguistic database for exploring information structure. In *Interdisciplinary Studies on Information Structure*, ISIS Working papers of the SFB 632 (1), pages 245–279.

Stefanie Dipper. 2005. XML-based stand-off representation and exploitation of multi-level linguistic annotation. In Rainer Eckstein and Robert Tolksdorf, editors, *Proceedings of Berliner XML Tage*, pages 39–50.

James B. Freeman. 1991. *Dialectics and the Macrostructure of Argument*. Foris, Berlin.

Michael Götze, Cornelia Endriss, Stefan Hinterwimmer, Ines Fiedler, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas, Ruben Stoel, and Thomas Weskott. 2007. Information structure. In *Information structure in cross-linguistic corpora: annotation guidelines for morphology, syntax, semantics, and information structure*, volume 7 of *ISIS Working papers of the SFB 632*, pages 145–187.

Olga Krasavina and Christian Chiarcos. 2007. Potsdam Coreference Scheme. In *this volume*.

Alan Lee, Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, and Bonnie Webber. 2006. Complexity of dependencies in discourse. In *Proc. 5^{th} Workshop on Treebanks and Linguistic Theory (TLT'06)*, Prague.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, 8:243–281.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large scale annotated corpus of English: The Penn TreeBank. *Computational Linguistics*, 19:313–330.

James R. Martin. 1992. *English text: system and structure*. John Benjamins, Philadelphia/Amsterdam.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. Annotating discourse connectives and their arguments. In *NAACL/HLT Workshop on Frontiers in Corpus Annotation*, Boston.

Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Experiments on sense annotation and sense disambiguation of discourse connectives. In *4^t Workshop on Treebanks and Linguistic Theory (TLT'05)*, Barcelona, Spain.

Michael J. Moravcsik and Poovanalingan Murugesan. 1975. Some results on the function and quality of citations. *Soc. Stud. Sci.*, 5:88–91.

The PDTB-Group. 2006. The Penn Discourse TreeBank 1.0 annotation manual. Technical Report IRCS 06-01, University of Pennsylvania.

Rashmi Prasad, Eleni Miltsakaki, Aravind Joshi, and Bonnie Webber. 2004. Annotation and data mining of the Penn Discourse TreeBank. In *ACL Workshop on Discourse Annotation*, Barcelona, Spain, July.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. 2007. Attribution and its annotation in the Penn Discourse TreeBank. *TAL (Traitement Automatique des Langues*.

Randolph Quirk, Sidney Greenbaum, Geoffry Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language.* Longman, New York.

Manfred Stede. 2004. The Potsdam commentary corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 96–102, Barcelona.

Manfred Stede. 2007. RST revisited: disentangling nuclearity. In Cathrine Fabricius-Hansen and Wiebke Ramm, editors, *'Subordination' versus 'coordination' in sentence and text – from a cross-linguistic perspective*. John Benjamins, Amsterdam. (to appear).

Simone Teufel and Marc Moens. 2002. Summarising scientific articles — experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–446.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the 9th European Conference of the ACL (EACL-99)*, pages 110–117, Bergen, Norway.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. An annotation scheme for citation function. In *Proceedings of SIGDIAL-06*, Sydney, Australia.

Simone Teufel. 2000. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, School of Cognitive Science, University of Edinburgh, Edinburgh, UK.

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

Bonnie Webber. 2005. A short introduction to the Penn Discourse TreeBank. In *Copenhagen Working Papers in Language and Speech Processing*.

Janyce Wiebe, Eric Breck, Chris Buckley, Claire Cardie, Paul Davis, Bruce Fraser, Diane Litman, David Pierce, Ellen Riloff, Theresa Wilson, David Day, and Mark Maybury. 2003. Recognizing and organizing opinions expressed in the world press. In *Working Notes of the AAAI Spring Symposium in New Directions in Question Answering*, pages 12–19, Palo Alto, California.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2/3):164–210.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31:249–287.