The Penn Discourse TreeBank as a Resource for Natural Language Generation

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki
Institute for Research in Cognitive Science
University of Pennsylvania

Bonnie Webber
Division of Informatics
University of Edinburgh

Workshop on Using Corpora for NLG
Birmingham, U.K.
July 14, 2005

Introduction

- Many NLG systems include a representation of discourse relations (DR) in the text plan.
- Lacking robust, reliable theories, NLG must rely on corpus analysis to suggest preferences in and constraints on DR realization.
- **GOAL:** To show how the Penn Discourse TreeBank (PDTB), which encodes explicitly discourse connectives and their arguments, can be used as a large scale annotated corpus resource for several NLG tasks:

  I. The sentence planning task of DR lexicalization: Occurrence, Selection, Placement.

  II. The text planning task of representing Attribution in the News report domain.

  III. The sentence planning task of realizing Attribution in this domain.

N.B. This is not a claim that discourse connectives are the only way to realize discourse relations.

Outline

- Overview of annotations in the PDTB.
- PDTB and DR-lexicalization during sentence planning:
  - Occurrence.
  - Selection.
  - Placement.
- PDTB and the representation of attribution.
- PDTB and the realization of attribution.
- Conclusion and Summary.

The Penn Discourse TreeBank: an Overview

- The PDTB contains annotations of discourse relations anchored on discourse connectives.
- Annotation approach is grounded in a lexicalized approach to low-level discourse structure (Webber *et al.* [2003]).
- Annotated discourse relations are semantic relations holding between TWO abstract objects (Asher [1993]):

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

- Arguments are named Arg1 and Arg2:
  Arg2 is the argument syntactically bound to the connective.
- Types of Annotations:
  - Explicit discourse connectives and their arguments.
  - Implicit discourse connectives and their arguments.
  - Semantic sense of each discourse connective.
  - Semantic (abstract object) classification of each argument (Future Work).
  - Attribution of discourse connectives and their arguments.

Explicit Connectives

- Explicit connectives include:
  - Subordinating conjunctions (*because*, *since*, *although*, *only because*, etc.)
  - Subordinators (*in order that*, *except that*, etc.)
  - Coordinating conjunctions (*and*, *or*, etc.)
  - Discourse adverbials (*however*, *otherwise*, etc.)

- For subordinating conjunctions, possible relative order of arguments:

  (2) Arg1-Arg2:
  Third-quarter sales in Europe were exceptionally strong, boosted by promotional programs and new products – **although** weaker foreign currencies reduced the company's earnings.

  (3) Arg2-Arg1:
  although she drives a canary-colored Porsche, she hasn't time to clean or repair it.

  (4) Arg1-Arg2-Arg1:
  Most oil companies, **when** they set exploration and production budgets for this year, forecast revenue of $15 for each barrel of crude produced.

- For subordinators and coordinating conjunctions, argument order is Arg1-Arg2.

- For discourse adverbials, typical argument order Arg1-Arg2, but embedding possible:

  (5) Arg2-Arg1-Arg2:
  market analysts said that late Tuesday the Chinese government, which often buys U.S. grains in quantity, turned **instead** to Britain to buy 500,000 metric tons of wheat.

Implicit Connectives

- Annotated between adjacent sentences when no connective appears explicitly to relate the second sentence to the first.

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

- Multiple "simultaneous" interpretations also annotated:

  (7) The small, wiry Mr. Morishita comes across as an outspoken man of the world. IMPLICIT=WHEN IMPLICIT=FOR EXAMPLE Stretching his arms in his silky white shirt and squeaking his black shoes he lectures a visitor about the way to sell American real estate and boasts about his friendship with Margaret Thatcher's son.

- Argument text spans also marked for implicit connectives, but the spans may be more or less than an entire sentence.

- Not annotated: Possible implicit connectives between sentences across a paragraph boundary and intra-sentential implicit connectives (e.g. *free adjuncts*).

Extent of Arguments

- Abstract object arguments can arise from the interpretation of:
  single clauses, multiple clauses, single sentences, multiple sentences.

- Arguments can also be non-clausal units that denote abstract objects:
  nominalizations and discourse deictics (*this*, *that*).

- However, a *Minimality Principle* constrains argument selection to the minimal amount of information necessary for the interpretation of the relation. Any other span perceived to be relevant (but not necessary) is labelled *supplementary* to the arguments:

  Sup1 for Arg1:

  (8) **Although** started in 1965, Wedtech didn't really get rolling until 1975 when Mr. Neuberger discovered the Federal government's Section 8 minority program.

  Sup2 for Arg2:

  (9) Bankers said warrants for Hong Kong stocks are attractive **because** they give foreign investors, wary of volatility in the colony's stock market, an opportunity to buy shares without taking too great a risk.

- Minimality Principle has analog at sentence level (Stone and Webber [1998]; Stone *et al.* [2001]).

Sense Annotation

- Each connective (explicit and implicit) will be annotated with a label corresponding to the sense of the projected discourse relation.

- Of particular interest are polysemous connectives.

- Preliminary senses for *since*:
  - Purely TEMPORAL:
    (10) The Mountain View, Calif., company has been receiving 1,000 calls a day about the product **since** it was demonstrated at a computer publishing conference several weeks ago.
  - Purely CAUSAL:
    (11) It was a far safer deal for lenders **since** NWA had a healthier cash flow and more collateral on hand.
  - Both CAUSAL and TEMPORAL:
    (12) . . . and domestic car sales have plunged 19% **since** the Big Three ended many of their programs Sept. 30.

Attribution Annotation

- *Attribution:* ascribing beliefs and assertions expressed in text to the agent(s) holding or making them.

- Two sources of attribution:
  Writer ("Writer attribution") or **Speaker** ("Speaker attribution").

- **Case 1:** Relation and both arguments are attributed to the same source:

  (13) "The public is buying the market **when** in reality there is plenty of grain to be shipped", said Bill Biedermann, Allendale Inc. research director.

- **Case 2:** One or both arguments have a different attribution value from the relation:

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

- Further refinement of attribution annotation underway:
  - distinctions between verbs of saying and verbs of propositional attitude.
  - factuality.
  - negation.

I. PDTB and Sentence Planning

- *Assumed architecture for NLG systems:* Pipeline, with *sentence planning* as an intermediate component between *text planning* and *surface realization* (Rambow and Korelsky [1992]).

- Input to sentence planning assumed to be a hierarchically ordered *text plan structure*:
  - Leaves ≈ elementary content units (ECUs)
  - Internal nodes ≈ discourse relations between ECUs or groups of ECUs.

- DR-lexicalization subsumes *aggregation* and includes a set of three lexicalization tasks (Moser and Moore [1995]):
  - *occurrence:* whether or not to generate an explicit connective
  - *selection:* which connective to generate
  - *placement:* where to place the generated connective.

Related work on DR-lexicalization

- Work to date on DR-lexicalization in generation has been somewhat spotty:
  - It has singled out a few connectives, e.g., (Elhadad and McKeown [1990]).
  - Heuristics have been based on a small number of constructed examples, e.g., (Scott and Souza, 1990).
  - Classification-based lexicons have been based on individual languages and may be manually-intensive to build in a multilingual context, e.g., (Grote and Stede [1998]; Knott and Mellish [1996]).
- Corpora annotated with information about connectives can provide a valuable resource for research on discourse connectives and discourse relations, or for inducing lexicons and models for generation of discourse connectives.

Occurrence

- Are there constraints on how a relation is lexicalized?

  (15) The three men worked together on the so-called Brady Commission, headed by Mr. Brady, which was established after the 1987 crash to examine the market's collapse. **As a result** they have extensive knowledge in fi nancial markets, and fi nancial market crises.

  (16) From 1984 to 1987, its (Iverson's) earnings soared six-fold, to $3.8 million, on a seven-fold increase in revenue, to $44.1 million. But in 1988, it ran into a buzz saw: a Defense Department spending freeze. IMPLICIT=AS A RESULT Iverson's earnings plunged 70% to $1.2 million.

- In some cases, yes:
  - (Amsili and Rossari [1998]): In French, the realization of a CAUSAL relation between eventualities is constrained by their *aspectual classes* and the *order* in which the eventualities appear in the relation.
  - (Williams and Reiter [2003]): There are statistically significant differences across classes of connectives with respect to how frequently they are lexicalized.

- Elsewhere we don't yet know: PDTB annotations on both explicit and implicit connectives will highlight where to look for constraints on the choice for lexicalization.

Learning from Absence or Gaps in Lexicalizability

- For learning about occurrence, it is useful that PDTB annotators can indicate where:
  - They can infer no discourse relation (NOREL):
    (17) The transaction has been approved by Kyle's board, but requires the approval of the company's shareholders. IMPLICIT=NOREL Kyle manufactures electronic components.
  - The inferred relation ≈ entity elaboration (NOCONN-ENT):
    (18) C.B. Rogers Jr. was named chief executive officer of this business information concern. IMPLICIT=NOCONN-ENT Mr. Rogers, 60 years old, succeeds J.V. White, 64, who will remain chairman and chairman of the executive committee.
  - They found unacceptable the result of lexicalizing the inferred relation as a connective (NOCONN):
    (19) In the 1920s, a young schoolteacher, John T. Scopes, volunteered to be a guinea pig in a test case sponsored by the American Civil Liberties Union to challenge a ban on the teaching of evolution imposed by the Tennessee Legislature. IMPLICIT-NOCONN The result was a world-famous trial exposing profound cultural conflicts in American life between the "smart set," . . . and the religious fundamentalists, . . .

Selection

- What constrains the choice of a particular connective, if the relation is to be lexicalized?
- PDTB anntotations are anchored on discourse connectives, allowing inferences to be drawn more easily about their behavior, and claims made about them to be tested.
- **Claim 1 –** *because* and CAUSAL *since*:
  - (Elhadad and McKeown [1990]): Clause order correlates with the information status of the arguments:
    - *Because* tends to place Arg2 at the end, associating it with *new* information.
    - *since* tends to place Arg2 at the beginning, associating it with *given* information.
  - In the PDTB,
    - *because* does tend to appear postposed - 90% (Prasad *et al.* [2004])
    - but the 90 confirmed instances of CAUSAL *since* are equally distributed in pre- and postposed positions.
  - Further clarification thus needed for the correlation between clause order and information status.

- **Claim 2 –** *although* and *even though*:
  - Assumptions (Huddleston and Pullum [2002]):
    - “Even” reinforces concessive meaning.
    - “Although” and “though” are themselves indistinguishable.
  - But in the PDTB, they are complementary in the relative position of their arguments.

CONN	*Arg2* Postposed	*Arg2* Preposed	**Total**
although	129 (37%)	218 (63%)	347
even though	77 (75%)	26 (25%)	103
though	97 (70%)	42 (30%)	139
Total	303 (51%)	286 (49%)	589

  - It may be that *even though* simply patterns like *though*, and can be grouped together.
  - Alternatively, *even though* may have an underlying bi-modal distribution, part from *though+even*, and part from *although+even*, since the latter is also realized as *even though*. Clearly more analysis is required.

Placement

- When relating two CUs, the discourse connective to be lexicalized is syntactically bound to one of the CUs (Arg2).

- **Task 1:** Which CU to associate the connective with? (Which CU will be Arg2?)
  - Fairly simple resolution: assuming linear ordering to be given, Arg2 can be determined by:
    - semantics of the particular relation viz-a-viz the linear order of the arguments
    - syntactic type of the connective
  - Example: semantics=denial of expectation
    - If CONN is subordinating conjunction, associate with CU that raises the expectation.
      (20) **Although** John is smart, he failed the exam.
      (21) John failed the exam **even though** he is smart.
    - If CONN is coordinating conjunction, associate with CU that denies the expectation.
      (22) John failed the exam **but** he is smart.

- **Task 2:** For discourse adverbials, where to place the connective in its Arg2 CU?

  – Connective can appear in *initial*, *medial*, or *final* position.

  (23) Despite the economic slowdown, there are few clear signs that growth is coming to a halt. **As a result**, Fed officials may be divided over whether to ease credit.

  (24) The chief culprits, he says, are big companies and business groups that buy huge amounts of land "not for their corporate use, but for resale at huge profit." . . . The Ministry of Finance, **as a result**, has proposed a series of measures that would restrict business investment in real estate . . .

  (25) Polyvinyl chloride capacity "has overtaken demand and we are experiencing reduced profit margins **as a result**.", . . .

  – Early experiments on PDTB (Prasad *et al.* [2004]) looked at the position of 5 adverbials:

    * *as a result*, *instead*, *nevertheless*, *otherwise*, *therefore*.

    * connectives in this group occurred predominantly in *initial* position.

    * Further experiments are needed to determine behavior of other discourse adverbials.

Benefits of PDTB for Inducing Corpus-based models for Sentence Planning

- With the PDTB, the benefits of corpus-induced constraints are greater than those obtained from previous corpus-based studies and experiments:
  - Size of the corpus is much larger (1M words).
  - PDTB is aligned with other layers of annotation: syntactic annotation of the Penn TreeBank and semantic annotation of the PropBank.
  - For NLG systems that assume the syntactic and semantic specification of CUs to be given for sentence planning in the text plan, the PDTB will allow induction of a much richer set of constraints for sentence planning.
  - Some work on automatically inducing models for the selection task has been done: e.g., (Hutchinson [2005]), but these models exploit only *lexical cooccurrences*.
  - Keeping the generation architecture in perspective, such modeling will benefit more by carrying out similar experiments on the multi-layered annotation of the PDTB.

II. Recognizing Attribution in Domain Modeling

- Most working NLG systems are built within restricted domains.
- While some tasks (sentence planning, surface realization) are still modeled in a general way, *content determination* and *text planning* are driven by the needs of the domain.
- *domain modeling:* analyzing target texts and declaring the types of information that needs to be conveyed.
  - Restaurant review domain (Walker *et al.* [2003]): entities such as *restaurant*, with properties like *food*, *service*, *atmostphere* predicated over these entities.
  - Weather domain (Reiter and Dale [2000]): entities such as *time-span*, with properties such as *rainfall* defined over these entities.
- *News Domain:* more complex types of entities and relations.
  - Attribution: “ownership” relation between abstract objects and individuals or agents.
  - Beliefs and statements should be attributed to their correct sources (“owners”).
  - False inference should be avoided that some information is a commonly known fact or commonly held belief.

Two Types of Facts in a Report

- Factual status of “chief culprits are big companies and business groups”

  – Taken to be a fact from some individual’s point of view:

    (26) The chief culprits, he (Mr. Lee) says, are big companies and business groups that buy huge amounts of land “not for their corporate use, but for resale at huge profit.”

  – Taken to be a fact of common knowledge, and hence to be **true**:

    (27) The chief culprits are big companies and business groups that buy huge amounts of land “not for their corporate use, but for resale at huge profit.”

Attribution Defined over Discourse Relations

- Attribution can also be defined over discourse relations.

  (28) "**When** the airline information came through, it cracked every model we had for the marketplace," said a managing director at one of the largest program-trading firms.

- Further extension needed of the abstract object (AO) ontology and its representation by the text planner.

- PDTB annotations can serve as a useful target corpus for studying the different types of entities and properties associated with attribution.

Representing Attribution during Text Planning

- *Assumption:* all information represented in the text plan must be realized in one way or another.

- PDTB evidence: Attribution has both an "owner" and a scope, and both must be correctly represented in the text plan for appropriate realization.

- **Case 1:** A discourse relation may hold between the attributions themselves:

  (29) Advocates said the 90-cent-an-hour rise, to $4.25 an hour by April 1991, is too small for the working poor, **while** opponents argued that the increase will still hurt small business and cost many thousands of jobs.

- **Case 2:** A discourse relation may hold just between the AO arguments of the attribution:

  (30) **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.

III. Realizing Attribution during Sentence Planning

- Attribution can be realized in different ways, affecting sentence planning directly:
  - Realization in quoted speech (28) or indirect speech (29).
  - Non-lexicalized attribution, to be recovered from prior discourse: (31).
  - Lexical choice of type of attribution: (32) v. (31)

(31) "Now, Philip Morris Kraft General Foods' parent company is committed to the coffee business and to increased advertising for Maxwell House," says Dick Mayer, president of the General Foods USA division. "**Even though** brand loyalty is rather strong for coffee, we need advertising to maintain and strengthen it."

(32) Like other large Valley companies, Intel also noted that it has factories in several parts of the nation, **so that** a breakdown at one location shouldn't leave customers in a total pinch.

- Constraints for different choices can be automatically derived from the PDTB.

Conclusions and Summary

- NLG systems representing discourse relations for realization need corpus resources like the PDTB for research and for induction of models for DR-lexicalization.
- Sentence planning tasks of occurrence, selection, and placement can benefi t directly from the PDTB since the annotations are anchored on discourse connectives.
- PDTB can serve as a useful target corpus for studying the ways in which the attribution relation should be represented in the text plan.
- Constraints on attribution realization can be induced from the corpus.
- **Summary of Penn Discourse TreeBank:**
  - First Release: November 2005
  - 16K explicit connectives (over 90 types)
    - 6000 subordinating conjunctions
    - 5000 coordinating conjunctions
    - 5000 discourse adverbials
  - 20K implicit connectives
  - Miltsakaki *et al.* [2004]; Prasad *et al.* [2004]; Webber *et al.* [2005]; Dinesh *et al.* [2005]
  - http://www.cis.upenn.edu/ pdtb
  - **N.B.** Slides for this talk are posted on the project webpage.

References

Pascal Amsili and Corinne Rossari. Tense and connective constraints on the expression of causality. In *Proc. COLING-ACL*, pages 48–54, 1998.

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

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

Michael Elhadad and Kathleen R. McKeown. Generating connectives. In *Proc. COLING*, volume 3, pages 97–101, 1990.

Brigitte Grote and Manfred Stede. Discourse marker choice in sentence planning. In *Proc. INLG*, pages 128–137, 1998.

Ronald Huddleston and Geoffrey Pullum. *The Cambridge Grammar of the English Language*. Cambridge Univ. Press, Cambridge, UK, 2002.

Ben Hutchinson. Modeling the substitutability of discourse connectives. In *Proc. ACL*, 2005.

Alistair Knott and Chris Mellish. A feature-based account of the relations signalled by sentence and clause connectives. *Language and Speech*, 39(2-3):143–183, 1996.

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

Megan G. Moser and Johanna D. Moore. Using discourse analysis and automatic text generation to the study of cue usage. In *Proc. AAAI Symposium on Empirical Methods in Discourse Interpretation and Organization*, pages 92–98, 1995.

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

Owen Rambow and Tanya Korelsky. Applied text generation. In *Proc. ANLP*, pages 40–47, 1992.

Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Cambridge Univ. Press, 2000.

Matthew Stone and Bonnie Webber. Textual economy through close coupling of syntax and semantics. In *Proc. INLG*, pages 178–187, 1998.

Matthew Stone, Christine Doran, Bonnie Webber, Tonia Bleam, and Martha Palmer. Microplanning from communicative intentions: Sentence planning using descriptions (spud). *Computational Intelligence*, 2001.

Marilyn Walker, Rashmi Prasad, and Amanda Stent. A trainable generator for recommendations in multimodal dialogue. In *Proc. EUROSPEECH*, pages 1697–1701, 2003.

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

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

Sandra Williams and Ehud Reiter. A corpus analysis of discourse relations for natural language generation. In *Proc. Corpus Linguistics*, pages 899–908, 2003.