Explicit discourse connective argument identification with connective specific rankers, with additional thoughts

Jason Baldridge
Department of Linguistics
University of Texas at Austin
Explicit discourse connective argument identification

- Full discourse structure is difficult and there is (still) no agreed upon representation. However, many discourse relations are signaled by explicit discourse connectives, like *because* and *then*.

- Task: identify the textual arguments of explicit discourse connectives.
Explicit discourse connective argument identification

- Full discourse structure is difficult and there is (still) no agreed upon representation. However, many discourse relations are signaled by explicit discourse connectives, like *because* and *then*.

- Task: identify the textual arguments of explicit discourse connectives.

Max fell because John pushed him.
Explicit discourse connective argument identification

- Full discourse structure is difficult and there is (still) no agreed upon representation. However, many discourse relations are signaled by explicit discourse connectives, like *because* and *then*.

- Task: identify the textual arguments of explicit discourse connectives.

Max fell **because** John pushed him.

discourse connective
Explicit discourse connective argument identification

- Full discourse structure is difficult and there is (still) no agreed upon representation. However, many discourse relations are signaled by explicit discourse connectives, like *because* and *then*.

- Task: identify the textual arguments of explicit discourse connectives.

```
```

```
Arg1
```

discourse connective
Explicit discourse connective argument identification

- Full discourse structure is difficult and there is (still) no agreed upon representation. However, many discourse relations are signaled by explicit discourse connectives, like *because* and *then*.

- Task: identify the textual arguments of explicit discourse connectives.

```
Max fell because John pushed him.
```
Explicit discourse connective argument identification

- Full discourse structure is difficult and there is (still) no agreed upon representation. However, many discourse relations are signaled by explicit discourse connectives, like *because* and *then*.

- Task: identify the textual arguments of explicit discourse connectives.

```
Max fell because John pushed him.
```

- **Arg1**: Max (head)
- **Arg2**: John (head)
- **Discourse connective**: because
Big picture: approach

- Given training material this takes the form of a straightforward classification task.

- Previous work by Wellner and Pustejovsky (2007) used a single supervised ranking model.

- Elwell and Baldridge (2008) provided two enhancements:
  - **Model**: Interpolated models based on specific connectives and types of connectives.
  - **Features**: greater sensitivity to patterns of morphology, syntax, and discourse.
Discourse connectives

- Discourse connectives are explicit markers of rhetorical relations.

- There are several types:
  - Coordinating: *and, or, but, yet, then*
  - Subordinating: *because, when, since, even though, except when*
  - Adverbial: *afterwards, previously, nonetheless, actually*

- Connectives can be related to spans of text representing the arguments of the rhetorical relation they mark.
Penn Discourse Treebank (PDTB)

- The PDTB annotates the arguments of discourse connectives in Wall Street Journal texts (same as Penn Treebank).

- Retains a close link with the data: annotations are made directly on text spans.

- No abstract relations: only connectives themselves are annotated. (Some implicit relations are marked, but we ignore them here.)

- No higher level structure: avoids structural ambiguity that arises with efforts based on Rhetorical Structure Theory and Segmented Discourse Representation Theory.
Coordinating connective

Choose 203 business executives, including, perhaps, someone from your own staff, and put them out on the streets, to be deprived for one month of their homes, families, and income.

Subordinating connective

Drug makers shouldn't be able to duck liability because people couldn't identify precisely which identical drug was used.

Adverbial connective

France's second-largest government-owned insurance company, Assurances Generales de France, is building its own Navigation Mixte stake, currently thought to be between 8% and 10%. Analysts said they don't think it is contemplating a takeover, however, and its officials couldn't be reached.
Corpus examples

Long-distance subordinating connective

But while International Business Machines Corp. and Compaq Computer Corp. say the bugs will delay products, *most big computer makers said the flaws don't affect them.*``Bugs like this are just a normal part of product development," said Richard Archuleta, director of Hewlett-Packard Co.'s advanced systems development. Hewlett announced last week that it planned to ship a computer based on the 486 chip early next year. “These bugs don't affect our schedule at all,” he said. “Likewise, AST Research Inc. and Sun Microsystems Inc. said the bugs won't delay their development of 486-based machines. “We haven't modified our schedules in any way,” said a Sun spokesman. To switch to another vendor's chips, “would definitely not be an option," he said. Nonetheless, *concern about the chip may have been responsible for a decline of 87.5 cents in Intel's stock to $32 a share yesterday in over-the-counter trading, on volume of 3,609,800 shares, and partly responsible for a drop in Compaq's stock in New York Stock Exchange composite trading on Wednesday.*
But while International Business Machines Corp. and Compaq Computer Corp. say the bugs will delay products, most big computer makers said the flaws don't affect them. "Bugs like this are just a normal part of product development," said Richard Archuleta, director of Hewlett-Packard Co.'s advanced systems development. Hewlett announced last week that it planned to ship a computer based on the 486 chip early next year. "These bugs don't affect our schedule at all," he said. "Likewise, AST Research Inc. and Sun Microsystems Inc. said the bugs won't delay their development of 486-based machines. "We haven't modified our schedules in any way," said a Sun spokesman. To switch to another vendor's chips, "would definitely not be an option," he said. Nonetheless, concern about the chip may have been responsible for a decline of 87.5 cents in Intel's stock to $32 a share yesterday in over-the-counter trading, on volume of 3,609,800 shares, and partly responsible for a drop in Compaq's stock in New York Stock Exchange composite trading on Wednesday.
But while International Business Machines Corp. and Compaq Computer Corp. say the bugs will delay products, most big computer makers said the flaws don't affect them. "Bugs like this are just a normal part of product development," said Richard Archuleta, director of Hewlett-Packard Co.'s advanced systems development. Hewlett announced last week that it planned to ship a computer based on the 486 chip early next year. "These bugs don't affect our schedule at all," he said. "Likewise, AST Research Inc. and Sun Microsystems Inc. said the bugs won't delay their development of 486-based machines. "We haven't modified our schedules in any way," said a Sun spokesman. To switch to another vendor's chips, "would definitely not be an option," he said. Nonetheless, concern about the chip may have been responsible for a decline of 87.5 cents in Intel's stock to $32 a share yesterday in over-the-counter trading, on volume of 3,609,800 shares, and partly responsible for a drop in Compaq's stock in New York Stock Exchange composite trading on Wednesday.
Overlapping connectives

John loves Barolo. He ordered three cases of the '97. But he had to cancel the order because then he discovered he was broke.

John loves Barolo. He ordered three cases of the '97. But he had to cancel the order because then he discovered he was broke.
Recapping

- There are several different types of connectives, based on how they select ARG1s. Here, we consider coordinating, subordinating, and adverbial connectives.

- All connectives in PDTB find their ARG2s in their local syntactic context: ARG2s are structurally dependent.

- Connective arguments are annotated without respect to syntax.

- Syntactic analyses of Penn Treebank are nonetheless available during modeling.

- Annotations are non-recursive, so there is no overall graph spanning the entire text.
Automatic discourse connective argument identification

- First modeled by Wellner and Pustejovsky (2007).
- Given a connective, identify the text spans associated with its ARG1 and ARG2.
  - Simplifying assumption: identify the head word of the correct span
- Used maximum entropy ranker to identify best candidate.
- Reranked output for both ARG1 and ARG2 for joint prediction.
- Best results: ARG1: 76.4%  ARG2: 95.4%
France's second-largest government-owned insurance company, Assurances Generales de France, is building its own Navigation Mixte stake, currently thought to be between 8% and 10%. Analysts said they don't think it is contemplating a takeover, however, and its officials couldn't be reached.
Like Wellner and Pustejovksy, we use a maximum entropy ranker which compares all candidates simultaneously. So, we compute $\text{Prob}(\text{candidate} \mid \text{observations}, \text{candidates})$ rather than $\text{Prob}(\text{isArg1} \mid \text{observations})$.

This has worked well in other tasks, such as anaphora resolution, coreference, and question-answering.

We modify the features slightly, removing some of them to rely less on syntactic structures.

This model lumps all connectives together (albeit using features such as $\text{connective} = \text{because}$). We call this the General Connective model (GC).
Different types of connectives behave differently.

- coordinating and subordinating connectives generally find their ARG1 in the same or previous sentence
- adverbial connectives can find theirs several sentences back

Like coreference: the antecedents of pronouns are generally found nearby, unlike those of proper nouns.

Lumping all of them together means that we are attempting to capture different distributions with single model.

We create a model composed of three separate models for types of connectives (TC)
Likewise, particular connectives select their arguments differently, even within the same type.

We thus build a collection of models, one for each specific connective (SC).

This is similar to practice in other NLP tasks, such as word sense disambiguation.

Even part-of-speech taggers use word-specific models, in the form of tag dictionaries as hard filters on admissible tags for a word.
Model combination

- For SC and GC, there are fewer training instances per model, so they might suffer from sparsity despite having more a appropriate training distribution.

- We thus combine the models using simple interpolation:

\[ P_{TG}(a|c_i) = \lambda_{t_i} P_{t_i}(a) + (1 - \lambda_{t_i}) P_g(a) \]
\[ P_{SGT}(a|c_i) = \lambda_{c_i} P_{c_i}(a) + (1 - \lambda_{c_i}) P_{TG}(a|c_i) \]

- Interpolation weights are set such that rarer connectives or types leave more mass for the more general models.

\[ \lambda_{c_i} = \frac{freq(c_i)}{freq(c_i) + C} \]
France's second-largest government-owned insurance company, Assurances Generales de France, is building its own Navigation Mixte stake, currently thought to be between 8% and 10%. Analysts said they don't think it is contemplating a takeover, however, and its officials couldn't be reached.
France's second-largest government-owned insurance company, Assurances Generales de France, is building its own Navigation Mixte stake, currently thought to be between 8% and 10%. Analysts said they don't think it is contemplating a takeover, however, and its officials couldn't be reached.
France's second-largest government-owned insurance company, Assurances Generales de France, is building its own Navigation Mixte stake, currently thought to be between 8% and 10%. Analysts said they don't think it is contemplating a takeover, however, and its officials couldn't be reached.
France's second-largest government-owned insurance company, Assurances Generales de France, is building its own Navigation Mixte stake, currently thought to be between 8% and 10%. Analysts said they don't think it is contemplating a takeover, however, and its officials couldn't be reached.
France's second-largest government-owned insurance company, Assurances Generales de France, is building its own Navigation Mixte stake, currently thought to be between 8% and 10%. Analysts said they don't think it is contemplating a takeover, however, and its officials couldn't be reached.
France's second-largest government-owned insurance company, Assurances Generales de France, is building its own Navigation Mixte stake, currently thought to be between 8% and 10%. Analysts said they don't think it is contemplating a takeover, however, and its officials couldn't be reached.
France's second-largest government-owned insurance company, Assurances Generales de France, is building its own Navigation Mixte stake, currently thought to be between 8% and 10%. Analysts said they don't think it is contemplating a takeover, however, and its officials couldn't be reached.
France's second-largest government-owned insurance company, Assurances Generales de France, is building its own Navigation Mixte stake, currently thought to be between 8% and 10%. Analysts said they don't think it is contemplating a takeover, however, and its officials couldn't be reached.
Evaluation

- Standard train/dev/test split (PTB 2-22/0-1/23-24)
- Accuracy for:
  - ARG1 only
  - ARG2 only
  - CONN: both ARG1 and ARG2
- CONN performance broken down by connective type.
- Performance using features extracted from both gold parses and auto-parses (Bikel parser)
Results with gold-standard parses

- Arg1
- Arg2
- CONN

- W&P reranker
- GC
- TC
- SC
- GC-TC-SC
Results with gold-standard parses

Arg1 accuracy better than W&P reranker
Results with gold-standard parses

Arg1 accuracy better than W&P reranker

Arg2 accuracy worse than W&P reranker

<table>
<thead>
<tr>
<th></th>
<th>W&amp;P reranker</th>
<th>GC</th>
<th>TC</th>
<th>SC</th>
<th>GC-TC-SC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Monday, April 30, 12
Results with gold-standard parses

Arg1 accuracy better than W&P reranker

Arg2 accuracy worse than W&P reranker

Joint accuracy better than W&P reranker: 77.8 vs 74.2
Results with gold-standard parses

Arg1 accuracy of TC beats GC (better distribution) and SC (less sparsity).
Results with auto-parses (Bikel)

- Arg1
- Arg2
- CONN

- W&P reranker
- W&P reranker (gold)
- GC-TC-SC
- GC-TC-SC (gold)
Results with auto-parses (Bikel)

Drop due to auto-parses significant, but not drastic.
Results with auto-parses (Bikel)

Drop due to auto-parses significant, but not drastic.

Less impact than for W&P.
- fewer syntactic features
- used 5-fold generation of auto-parses (W&P didn’t (Wellner, p.c.))
CONN accuracy by connective type

- Subord
- Coord
- Adverbial

Connective types:
- GC
- TC
- SC

© 2012 Jason M Baldridge

Monday, April 30, 12
SC best for Subordinating and Coordinating: sparsity less an issue because ArgI’s are found in nearby context.
SC best for Subordinating and Coordinating: sparsity less an issue because Arg1’s are found in nearby context.

TC best for Adverbial: balances sparsity with capturing split in Arg1 selection behaviour.
The Penn Discourse Treebank provides a useful level of annotation for rhetorical structure (there is more in version 2.0 not discussed here, such as attribution).

With two simple strategies, we have improved performance over previous work:

- Connective or type specific models, combined by interpolation.
- Features that capture more morphological and discourse context.

Improvements in CONN accuracy:

- 3.6% using gold-standard parses
- 9.0% using auto-parses
Classification wrap-up

- Our improvements are orthogonal to W&P’s reranker, so could be combined for further improvements.

- Also, we clearly would improve performance by using their features for the ARG2 model.

- Some connectives are of multiple types, like “since”, which we ignore here and could be modeled better. This would be especially important for dealing with implicit connectives.

- This is a very limited sub-space of the overall endeavor of full discourse parsing, e.g. see the system described by Lin et al 2010.
Feature engineering

- Feature engineering for explicit and implicit discourse relations and argument identification highly limited by amount of available training data.
  - Classifying implicit discourse relations: entity relations don’t help over lexical baseline. [Louis, Joshi, Prasad and Nenkova, 2010]

- Sparsity of relevant features is likely the issue, so this area may be fertile grounds for semi-supervised learning.

- Measuring effect with current resources also limited, suggesting task-based evaluations may help determine efficacy.
Higher level tasks

- Summarization: hard to beat strong lexical benchmarks. [Louis, Joshi, and Nenkova, 2010]

- Coreference: negative result on using SDRT structures as features for coreference resolution - unhelpful for the standard MUC coreference problem. More potential for phenomena like bridging.

- Areas where prediction of the rhetorical relations between statements might help more?
  - Identifying logical fallacies: do certain fallacies have characteristic rhetorical sequences or differential use of explicit vs implicit relations?
  - Detecting bias: does an author reveal their bias by the patterns of how statements are connected to rhetorical relations?
Different data: not(Wall Street Journal)

- **Social media**
  - branching discussions, with explicit indications via @-mentions
  - sentiment analysis

- **Contentious Wikipedia pages**
  - the record of edits and discussion histories could be used as latent indicators of rhetorical strategies and choices.
  - differential rates of footnotes
  - *_controvers(ylies)

Wikipedia deletion discussions
[http://notabilia.net/](http://notabilia.net/)
Different data: not(Wall Street Journal)

- Political speeches and debates
  - tied to voting records (implicit evidence of stances and likely rhetorical strategies)
  - measuring control and influence of moderators and debaters
Different data: not(Wall Street Journal)

- Real-time polling and reactions: compare time-aligned annotations of spin and approval/disapproval to rhetorical strategies employed.

- Philip Resnik’s ReactLabs

GOP DEBATE FEB 22 V2. February 22, 2012, 07:45pm EST.
Thanks!

- Joint work with
  - Nicholas Asher
  - Pascal Denis
  - Robert Elwell
  - Julie Hunter
  - Elias Ponvert
  - Brian Reese
  - Ben Wing

- This work was funded by NSF grant IIS-0535154.
<table>
<thead>
<tr>
<th>Model</th>
<th>Arg1</th>
<th>Arg2</th>
<th>Conn</th>
</tr>
</thead>
<tbody>
<tr>
<td>W&amp;P-Base</td>
<td>75.0</td>
<td>94.2</td>
<td>71.7</td>
</tr>
<tr>
<td>W&amp;P-Rerank</td>
<td>76.4</td>
<td>95.4</td>
<td>74.2</td>
</tr>
<tr>
<td>GC-W&amp;PFeats</td>
<td>78.1</td>
<td>91.8</td>
<td>73.0</td>
</tr>
<tr>
<td>GC-AllFeats</td>
<td>78.7</td>
<td>92.1</td>
<td>73.9</td>
</tr>
<tr>
<td>TC-AllFeats</td>
<td>81.7</td>
<td>92.6</td>
<td>76.1</td>
</tr>
<tr>
<td>SC-AllFeats</td>
<td>80.3</td>
<td>93.1</td>
<td>75.8</td>
</tr>
<tr>
<td>TC-GC-Interp</td>
<td>81.7</td>
<td>93.2</td>
<td>77.2</td>
</tr>
<tr>
<td>SC-TC-GC-Interp</td>
<td>82.0</td>
<td>93.7</td>
<td>77.8</td>
</tr>
</tbody>
</table>
## Results - Parsed Data

<table>
<thead>
<tr>
<th>Model</th>
<th>Arg1</th>
<th>Arg2</th>
<th>Conn</th>
</tr>
</thead>
<tbody>
<tr>
<td>W&amp;P-Base</td>
<td>67.9</td>
<td>90.6</td>
<td>62.7</td>
</tr>
<tr>
<td>W&amp;P-Rerank</td>
<td>69.8</td>
<td>90.8</td>
<td>64.6</td>
</tr>
<tr>
<td>GC-W&amp;PFeats</td>
<td>76.9</td>
<td>89.0</td>
<td>70.2</td>
</tr>
<tr>
<td>GC-AllFeats</td>
<td>77.1</td>
<td>89.1</td>
<td>70.2</td>
</tr>
<tr>
<td>TC-AllFeats</td>
<td>78.9</td>
<td>89.7</td>
<td>72.4</td>
</tr>
<tr>
<td>SC-AllFeats</td>
<td>78.7</td>
<td>85.4</td>
<td>68.4</td>
</tr>
<tr>
<td>TC-GC-Interp</td>
<td>79.8</td>
<td>89.9</td>
<td>73.1</td>
</tr>
<tr>
<td>SC-TC-GC-Interp</td>
<td>80.0</td>
<td>90.2</td>
<td>73.6</td>
</tr>
</tbody>
</table>
## Connective Results - Gold

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GC-W&amp;PFeats</td>
<td>80.8</td>
<td>74.2</td>
<td>58.7</td>
</tr>
<tr>
<td>GC-AllFeats</td>
<td>80.9</td>
<td>75.5</td>
<td>59.8</td>
</tr>
<tr>
<td>TC-AllFeats</td>
<td>82.8</td>
<td>77.5</td>
<td>67.5</td>
</tr>
<tr>
<td>SC-AllFeats</td>
<td>83.9</td>
<td>78.0</td>
<td>59.0</td>
</tr>
<tr>
<td>TC-GC-Interp</td>
<td>82.8</td>
<td>77.5</td>
<td>67.8</td>
</tr>
<tr>
<td>SC-TC-GC-Interp</td>
<td>83.9</td>
<td>78.1</td>
<td>67.5</td>
</tr>
</tbody>
</table>
## Connective Results - Parsed

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GC-W&amp;PFeats</td>
<td>76.2</td>
<td>73.3</td>
<td>55.4</td>
</tr>
<tr>
<td>GC-AllFeats</td>
<td>76.9</td>
<td>73.5</td>
<td>54.0</td>
</tr>
<tr>
<td>TC-AllFeats</td>
<td>78.2</td>
<td>75.4</td>
<td>58.2</td>
</tr>
<tr>
<td>SC-AllFeats</td>
<td>67.9</td>
<td>77.4</td>
<td>53.2</td>
</tr>
<tr>
<td>TC-GC-Interp</td>
<td>77.3</td>
<td>76.5</td>
<td>60.7</td>
</tr>
<tr>
<td>SC-TC-GC-Interp</td>
<td>78.0</td>
<td>77.1</td>
<td>60.7</td>
</tr>
</tbody>
</table>