Semantics and Pragmatics of NLP
Data Intensive Approaches to Discourse Interpretation

Alex Lascarides

School of Informatics
University of Edinburgh
Outline

1. Narrative Text
   - Corpora and annotation
   - Features for machine learning
   - Results

2. Dialogue
   - Corpora and annotation
   - Probabilistic Modelling
   - Results

3. Machine learning SDRSs

4. Unsupervised learning

Marcu (1999)

Rhetorical Parsing

- derives automatically the discourse structure of texts:
  - discourse segmentation as trees.
- approach relies on:
  - manual annotation;
  - theory of discourse structure (RST);
  - features for decision-tree learning
- given any text:
  - identifies rhetorical rels between text spans, resulting in a (global) discourse structure.
- useful for: text summarisation, information extraction, . . .
Annotation

Corpora:
- MUC7 corpus (30 stories);
- Brown corpus (30 scientific texts);
- Wall Street (30 editorials);

Coders:
- recognise *elementary discourse units* (*edus*);
- build discourse trees in the style of RST;
[Although discourse markers are ambiguous,\(^1\)]\([\text{one can use them to build discourse trees for unrestricted texts:}\(^2\)]\[\text{this will lead to many new applications in NLP.}\(^3\)]\)
Discourse Segmentation

**Task:** process each lexeme (word or punctuation mark) and decide whether it is:

- a sentence boundary (*sentence-break*);
- an *edu*-boundary (*edu-break*);
- a parenthetical unit (*begin-paren, end-paren*);
- a non-boundary (*non*).

**Approach:** Think of features that will predict classes, and then:

- Estimate features from annotated text;
- Use decision-tree learning to combine features and perform segmentation.
Discourse Segmentation

Features:

- local context:
  - POS-tags preceding and following lexeme (2 before, 2 after);
  - discourse markers (*because, and*);
  - abbreviations;

- global context:
  - discourse markers that introduce expectations (*on the one hand*);
  - commas or dashes before end of sentence;
  - verbs in unit of consideration.
Discourse Segmentation

Results:

<table>
<thead>
<tr>
<th>Corpus</th>
<th>B1 (%)</th>
<th>B2 (%)</th>
<th>DT (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC</td>
<td>91.28</td>
<td>93.1</td>
<td>96.24</td>
</tr>
<tr>
<td>WSJ</td>
<td>92.39</td>
<td>94.6</td>
<td>97.14</td>
</tr>
<tr>
<td>Brown</td>
<td>93.84</td>
<td>96.8</td>
<td>97.87</td>
</tr>
</tbody>
</table>

B1: defaults to *none*.

B2: defaults to *sentence-break* for every full-stop and *none* otherwise.

DT: decision tree classifier.
Discourse Structure

**Task:** determine rhetorical rels and construct discourse trees in the style of RST.

**Approach:**
- exploits RST trees created by annotators;
- map tree structure onto SHIFT/REDUCE operations;
- estimate features from operations.
- relies on RST’s notion of a nucleus and satellite:
  - **Nucleus:** the ‘most important’ argument to the rhetorical relation.
  - **Satellite:** the less important argument; could remove satellites and get a summary (in theory!)
Example of Mapping from Tree to Operations

{SHIFT 1; SHIFT 2; REDUCE-ATTRIBUTION-NS; SHIFT3; REDUCE-JOINT-NN; SHIFT 4; REDUCE-CONTRAST-SN}
Discourse Structure

Operations:
- 1 SHIFT operation;
- 3 REDUCE operations: RELATION-NS, RELATION-SN, RELATION-NN.

Rhetorical relations:
- taken from RST;
- 17 in total: CONTRAST, PURPOSE, EVIDENCE, EXAMPLE, ELABORATION, etc.
Features

- **structural**: rhetorical relations that link the immediate children of the link nodes;
- **lexico-syntactic**: discourse markers and their position;
- **operational**: last five operations;
- **semantic**: similarity between trees ($\approx$ bags-of-words).
### Results:

<table>
<thead>
<tr>
<th>Corpus</th>
<th>B3 (%)</th>
<th>B4 (%)</th>
<th>DT (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC</td>
<td>50.75</td>
<td>26.9</td>
<td>61.12</td>
</tr>
<tr>
<td>WSJ</td>
<td>50.34</td>
<td>27.3</td>
<td>61.65</td>
</tr>
<tr>
<td>Brown</td>
<td>50.18</td>
<td>28.1</td>
<td>61.81</td>
</tr>
</tbody>
</table>

B3: defaults to SHIFT.
B4: chooses SHIFT and REDUCE operations randomly.
DT: decision tree classifier.
## Breaking Down the Results

### Recognition of EDUs:

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Recall (%)</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC</td>
<td>75.4</td>
<td>96.9</td>
</tr>
<tr>
<td>WSJ</td>
<td>25.1</td>
<td>79.6</td>
</tr>
<tr>
<td>Brown</td>
<td>44.2</td>
<td>80.3</td>
</tr>
</tbody>
</table>

### Recognising Tree Structure:

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Recall (%)</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC</td>
<td>70.9</td>
<td>72.8</td>
</tr>
<tr>
<td>WSJ</td>
<td>40.1</td>
<td>66.3</td>
</tr>
<tr>
<td>Brown</td>
<td>44.7</td>
<td>59.1</td>
</tr>
</tbody>
</table>

### Results on Recognising Rhetorical Relations:

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Recall (%)</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC</td>
<td>38.4</td>
<td>45.3</td>
</tr>
<tr>
<td>WSJ</td>
<td>17.3</td>
<td>36.0</td>
</tr>
<tr>
<td>Brown</td>
<td>15.7</td>
<td>25.7</td>
</tr>
</tbody>
</table>
Summary

Pros:
- automatic discourse segmentation and construction of discourse structure;
- standard machine learning approach using decision-trees;

Cons:
- heavily relies on manual annotation;
- can only work for RST;
- no motivation for selected features;
- worst results on identification of rhetorical relations; but these convey information about meaning of text!
Dialogue Modelling


Automatic interpretation of dialogue acts:
- decide whether a given utterance is a question, statement, suggestion, etc.
- find the discourse structure of a conversation.

Approach relies on:
- manual annotation of conversational speech;
- a typology of dialogue acts;
- features for probabilistic learning;

Useful for: dialogue interpretation; HCI; speech recognition . . .
Dialogue Acts

A DA represents the meaning of an utterance at the level of illocutionary force (Austin 1962).
DAs ≈ speech acts (Searle 1969), conversational games (Power 1979).

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Dialogue Act</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>YES-NO-QUESTION</td>
<td>So do you go to college right now?</td>
</tr>
<tr>
<td>A</td>
<td>ABANDONED</td>
<td>Are yo-</td>
</tr>
<tr>
<td>B</td>
<td>YES-ANSWER</td>
<td>Yeah,</td>
</tr>
<tr>
<td>B</td>
<td>STATEMENT</td>
<td>It’s my last year [laughter].</td>
</tr>
<tr>
<td>A</td>
<td>DECL-QUESTION</td>
<td>So you’re a senior now.</td>
</tr>
<tr>
<td>B</td>
<td>YES-ANSWER</td>
<td>Yeah,</td>
</tr>
<tr>
<td>B</td>
<td>STATEMENT</td>
<td>I am trying to graduate.</td>
</tr>
<tr>
<td>A</td>
<td>APPRECIATION</td>
<td>That’s great.</td>
</tr>
</tbody>
</table>
**Corpus:** Switchboard, topic restricted telephone conversations between strangers (2430 American English conversations).

**Tagset:**
- DAMSL tagset (Core and Allen 1997);
- 42 tags;
- each utterance receives one DA (utterance $\approx$ sentence).
## Most Frequent DAs

<table>
<thead>
<tr>
<th>Category</th>
<th>Statement</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STATEMENT</strong></td>
<td><em>I’m in the legal department.</em></td>
<td>36%</td>
</tr>
<tr>
<td><strong>BACKCHANNEL</strong></td>
<td><em>Uh-huh.</em></td>
<td>19%</td>
</tr>
<tr>
<td><strong>OPINION</strong></td>
<td><em>I think it’s great.</em></td>
<td>13%</td>
</tr>
<tr>
<td><strong>ABANDONED</strong></td>
<td><em>So, -</em></td>
<td>6%</td>
</tr>
<tr>
<td><strong>AGREEMENT</strong></td>
<td><em>That’s exactly it.</em></td>
<td>5%</td>
</tr>
<tr>
<td><strong>APPRECIATION</strong></td>
<td><em>I can imagine.</em></td>
<td>2%</td>
</tr>
</tbody>
</table>
Automatic Classification of DAs

**Word Grammar:** Pick most likely DA given the word string (Gorin 1995, Hirschberg and Litman 1993), assuming words are independent:

\[ P(D|W) \]

**Discourse Grammar:** Pick most likely DA given surrounding speech acts (Jurafsky et al. 1997, Finke et al. 1997):

\[ P(D_i|D_{i-1}) \]

**Prosody:** pick most likely DA given acoustic ‘signature’ (e.g., contour, speaking rate etc.) (Taylor et al. 1996, Waibel 1998):

\[ P(D|F) \]
DA classification using Word Grammar

**Intuition:** utterances are distinguished by their words:
- 92.4% of *uh huhs* occur in BACKCHANNELS.
- 88.4% if *<s> do yous* occur in YES-NO-QUESTIONS.

**Approach:**
1. create a mini-corpus from all utterances which realise same DA;
2. train a separate word-\(N\)-gram model on each of these corpora.

**Task:** Given an utterance \(u\) consisting of word sequence \(W\), choose DA \(d\) whose \(N\)-gram grammar assigns highest likelihood to \(W\):

\[
d^* = \arg\max_d P(d|W) = \arg\max_d P(d)P(W|d)
\]
**Intuition:** the identity of previous DAs can be used to predict upcoming DAs.

**Task:** use $N$-gram models to model sequences of DAs. Dialogue act sequences are typically represented by HMMs.

- Bigram: $P(Yes|Yes-No-Question) = .30$
- Bigram: $P(Backchannel|Statement) = .23$
- Trigram: $P(Backchannel|Statement, Question) = .21$
A Dialogue Act HMM

YES-NO
QUESTION

YES

NO

STATEMENT

BCHANNEL

THANKING

YES

.76

.02

.23

.01

.62

.03

.22

.77

.46

.36

.18

.02

.18

.36

.46

.03

.22

.77

.03

.22

.77

.02

.18

.36

.46

.03

.22

.77
**Intuition:** prosody can help distinguish DAs with similar wordings but different stress.

- **STATEMENTS** pitch drops at the end.
- **YES-NO-QUESTIONS** pitch rises at the end.
- Without stress cannot distinguish **BACKCHANNEL**, **ANSWER-YES**, **AGREE**: all are often *yeah* or *uh-huh*.

**Prosodic Features:** duration, pauses, pitch, speaking rate, gender.

**Task:** build a decision-tree classifier that combines prosodic features to discriminate DAs.
Results

- 70.3% accuracy at detecting YES-NO-QUESTIONS;
- 75.5% accuracy at detecting ABANDONMENTS.
Combining Grammars

Given evidence $E$ about a conversation, find the DA sequence $\{d_1, d_2, \cdots, d_N\}$ with highest posterior probability $P(D|E)$.

$$D^* = \arg\max_D P(D|E) = \arg\max_D P(D)P(E|D)$$

Estimate $P(E|D)$ by combining word grammar $P(W|D)$ and prosody $P(F|D)$.

Choose DA sequence which maximises the product of conversational structure, prosody, and lexical knowledge.

$$D^* = \arg\max_D P(D)P(F|D)P(W|D)$$
## Results

<table>
<thead>
<tr>
<th>Discourse Grammar</th>
<th>Words</th>
<th>Prosody</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>42.8</td>
<td>38.9</td>
<td>56.5</td>
</tr>
<tr>
<td>Unigram</td>
<td>61.9</td>
<td>48.3</td>
<td>62.26</td>
</tr>
<tr>
<td>Bigram</td>
<td>64.6</td>
<td>50.2</td>
<td>65.0</td>
</tr>
</tbody>
</table>
Summary

Pros:
- automatic dialogue interpretation;
- standard probabilistic modelling;
- combination of different knowledge sources.

Cons:
- not portable between domains—manual annotation necessary;
- ignores non-linguistic factors:
  - relation between speakers, non-verbal behaviour,\ldots
- Not capturing hierarchical structure, so not useful for some (semantic) tasks.
Devise a (headed) tree representation from which SDRSs can be recovered:
- Leaves are utterances (marked with mood or ‘ignorable’ tag)
- Non-terminals are rhetorical relations, Segment or Pass.

Even though the representation is a tree, you can still recover SDRSs that aren’t trees:
- Pass node expresses $R_1(\alpha, \beta)$ and $R_2(\alpha, \gamma)$
- Node label as list of relations expresses $R_1(\alpha, \beta)$ and $R_2(\alpha \beta)$.

The heads determine which rhetorical relations have which arguments
Example

Tree:

```
  top
    └── dseg-h394-396
      └── dseg-h394-395
          └── contr
              └── dseg-h397-398
                  └── elabqp
                      └── qelab
                          └── pcorr/iqap
                              └── pelab
                                 └── pelab-h409-411
```

Relations Recovered from Tree:

```
pcorr(h397-398,h399-403), contr(h394-395,396), iqap(h397-398,h399-403), pcorr(393,h394-396), res*(h405-406,407),
elabqp(397,398), cont(401,402), expl*(407,408), cont(399,401), expl*(394,395), cont(405,406), expl(394,395), comnt(399,400),
pelab(h399-403,404), cont(409,411), cont(402,403), qelab(h394-396,h397-398), pelab(404,h405-408), pelab(h405-408,h409-411)
```
Learning A Discourse Parser

- Have annotated 100 dialogues with their discourse structure
- Because the representation is a tree, you can use standard sentential parsing models; we use Collins’ (1997) model.
- Features include things like:
  - Label of head daughter
  - Utterance tags
  - Number of speaker turns in the segment
  - The distance of the current modifier to the head daughter...
- Best model: 69% segmentation correct
  45% segmentation and rhetorical relations correct.
Pros and Cons

Pros:
- Allows one to use standard parsing techniques to build discourse structures that are hierarchical and *not* trees (cf. Marcu 1999).
- You get quite good results without recourse to rich features.
- Since SDRT has a model theory, you could use this discourse parser to automatically compute dialogue content, including implicatures.

Cons:
- Manual annotation is necessary; active learning might help.
- But it would be better to avoid annotating altogether!
Rhetorical relations can be overtly signalled: 

- *because* signals **EXPLANATION**; *but* signals **CONTRAST**

Use this to produce a training set *automatically*:

- Extract examples with unambiguous connectives; remove the connective and replace it with the relation it signals.
Marcu and Echihabi’s Model

It’s a Naive Bayes model using just word co-occurrences:

\[
P(r_i|W_1 \times W_2) = \frac{P(W_1 \times W_2|r_i)P(r_i)}{P(W_1 \times W_2)}
\] (1)

Since for any given example \(P(W_1 \times W_2)\) is fixed:

\[
\text{argmax } r_iP(r_i|W_1 \times W_2) = \text{argmax } r_iP(W_1 \times W_2|r_i)P(r_i)
\] (2)

With independence assumptions:

\[
P(W_1 \times W_2|r_i) \approx \prod_{(w_i, w_j) \in W_1 \times W_2} P((w_i, w_j)|r_i)
\] (3)

- Training set is very large: 9 million examples
- Achieves 48% accuracy on a six-way classifier.
Sporleder and Lascarides’ Model

Problem with Marcu and Echihabi:

- Smaller training sets sometimes necessary E.g., 8K examples of *in short* (for SUMMARY) on entire web!

Solution: More complex modelling and linguistic features

**Model:** Boostexter

**Features:** Verbs, verb classes, nouns, noun classes, adjectives
syntactic complexity, presence or absence of ellipsis
tense features, span length, positional features . . .

**Results:** Training set is 32K examples
Boostexter: 60.9%
Naive Bayes: 42.3%
Manually labelled 1K examples that *don’t* contain connectives with their rhetorical relation.

This is then used as the test set:
- Boostexter: 25.8%
- Naive Bayes: 25.9%

And as a training set:
- Boostexter: 40.3%
- Naive Bayes: 12%

So you’re better off manually labelling a small set of examples and using a sophisticated model!
Summary

Pros:
- No manual annotation of a training set is necessary

Cons:
- But it’s of limited use, because the resulting models perform poorly on examples that didn’t originally have a connective.
  - Lack of redundancy in the semantics of the clauses
  - Plurality of relations also a problem
Conclusions

Common features:

- approaches are corpus-based, and rely on:
  - annotation; feature extraction; probabilistic modelling.
- absence of symbolic reasoning;

Future Work:

- explore other ways of reducing manual annotation;
- explore different probabilistic models;
- apply models to unrestricted conversational speech, or to multi-agent dialogues
- combine probabilities with symbolic component;...