Semantics and Pragmatics of NLP
Data Intensive Approaches to Discourse Interpretation

Alex Lascarides

School of Informatics
University of Edinburgh

Outline

1. Narrative Text — Marcu (1999)
   - Corpora and annotation
   - Features for machine learning
   - Results
2. Dialogue — Stolcke *et al* (2000)
   - Corpora and annotation
   - Probabilistic Modelling
   - Results
3. Machine learning SDRSs
4. Unsupervised learning

Rhetorical Parsing

Marcu (1999)

- 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:

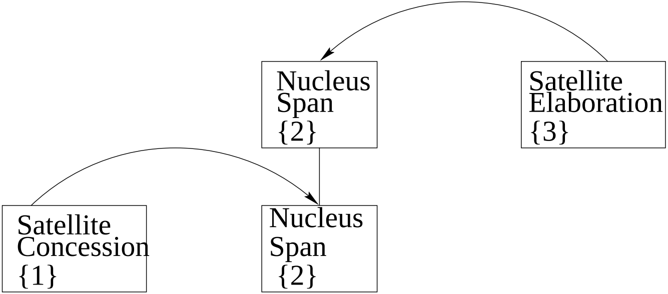
- MUC7 corpus (30 stories);
- Brown corpus (30 scientific texts);
- Wall Street (30 editorials);

Coders:

- recognise *elementary discourse units* (*edus*);
- build discourse trees in the style of RST;

Example

[Although discourse markers are ambiguous,[1]] [one can use them to build discourse trees for unrestricted texts:[2]] [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:

Corpus	B1 (%)	B2 (%)	DT (%)
MUC	91.28	93.1	96.24
WSJ	92.39	94.6	97.14
Brown	93.84	96.8	97.87

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.

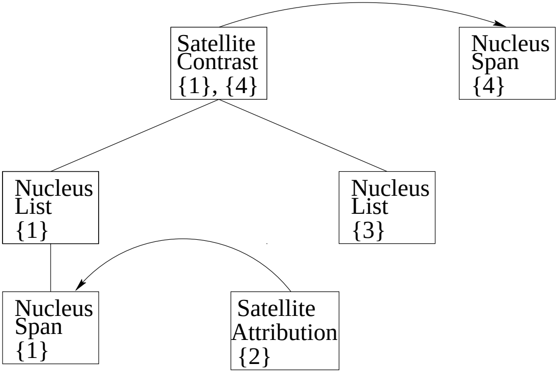
Approach:

- 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).

Discourse Structure

Results:

Corpus	B3 (%)	B4 (%)	DT (%)
MUC	50.75	26.9	61.12
WSJ	50.34	27.3	61.65
Brown	50.18	28.1	61.81

B3: defaults to SHIFT.
B4: chooses SHIFT and REDUCE operations randomly.
DT: decision tree classifier.

Breaking Down the Results

Recognition of EDUs:

Corpora	Recall (%)	Precision (%)
MUC	75.4	96.9
WSJ	25.1	79.6
Brown	44.2	80.3

Recognising Tree Structure:

Corpora	Recall (%)	Precision (%)
MUC	70.9	72.8
WSJ	40.1	66.3
Brown	44.7	59.1

Results on Recognising Rhetorical Relations:

Corpora	Recall (%)	Precision (%)
MUC	38.4	45.3
WSJ	17.3	36.0
Brown	15.7	25.7

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 — Stolcke *et al* (2000)

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).

Speaker	Dialogue Act	Utterance
A	YES-NO-QUESTION	*So do you go to college right now?*
A	ABANDONED	*Are yo-*
B	YES-ANSWER	*Yeah*,
B	STATEMENT	*It's my last year [laughter]*.
A	DECL-QUESTION	So you're a senior now.
B	YES-ANSWER	*Yeah*,
B	STATEMENT	*I am trying to graduate*.
A	APPRECIATION	That's great.

Annotation

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

STATEMENT	*I'm in the legal department.*	36%
BACKCHANNEL	*Uh-huh*.	19%
OPINION	*I think it's great.*	13%
ABANDONED	*So, -*	6%
AGREEMENT	*That's exactly it.*	5%
APPRECIATION	*I can imagine.*	2%

Automatic Classification of DAs

Word Grammar: Pick most likely DA given the word string (Gorin 1995, Hisrchberg 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 huh*s occur in BACKCHANNELS.
  88.4% if $<$s$>$ *do you*s 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. $P(W|d)$

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

$$d^* = \underset{d}{\operatorname{argmax}}\, P(d|W) = \underset{d}{\operatorname{argmax}}\, P(d)P(W|d)$$

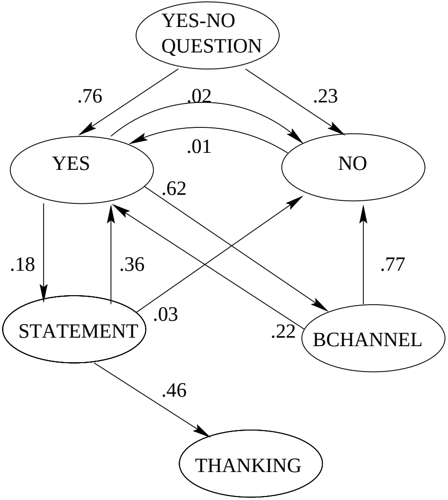
DA classification using Discourse Grammar

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:	$P(\mathit{Yes}\mid\mathit{Yes\text{-}No\text{-}Question}) = .30$
Bigram:	$P(\mathit{Backchannel}\mid\mathit{Statement}) = .23$
Trigram:	$P(\mathit{Backchannel}\mid\mathit{Statement},\mathit{Question}) = .21$

A Dialogue Act HMM



DA classification using Prosody

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 $P(D|E)$.

$$D^* = \operatorname*{argmax}_D P(D|E) = \operatorname*{argmax}_D P(D)P(E|D)$$

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.

$$D* = \operatorname*{argmax}_D P(D)P(F|D)P(W|D)$$

Results

Discourse Grammar	Words	Prosody	Combined
None	42.8	38.9	56.5
Unigram	61.9	48.3	62.26
Bigram	64.6	50.2	65.0

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,...
- Not capturing hierarchical structure, so not useful for some (semantic) tasks.

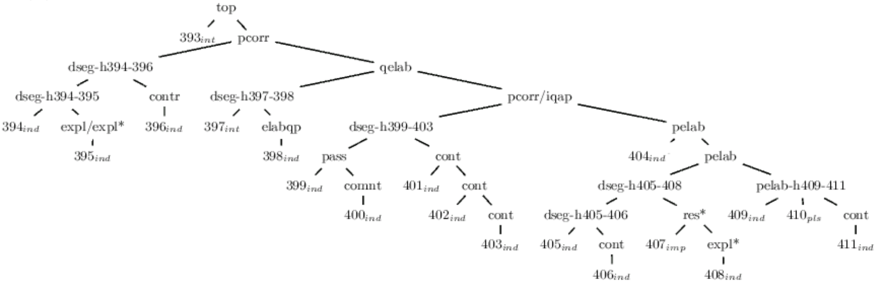
Building SDRSs for Dialogue

(Baldridge and Lascarides 2005)

- 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 reprsentation 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:



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!

Avoiding Annotation

Marcu and Echihabi 2002, Sporleder and Lascarides 2005

- 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:

$$P(r_i|W_1 \times W_2) = \frac{P(W_1 \times W_2|r_i)P(r_i)}{P(W_1 \times W_2)} \tag{1}$$

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

$$\text{argmax}\, r_i P(r_i|W_1 \times W_2) = \text{argmax}\, r_i P(W_1 \times W_2|r_i)P(r_i) \tag{2}$$

With independence assumptions:

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

- 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%

Both Perform Badly on Examples without Connectives!

- 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;...