Finding features to detect discourse relations between sentences within paragraphs

Daphne Theijssen
Master Thesis
Research Master Language and Communication
Radboud University Nijmegen
d.theijssen@gmail.com

This presentation

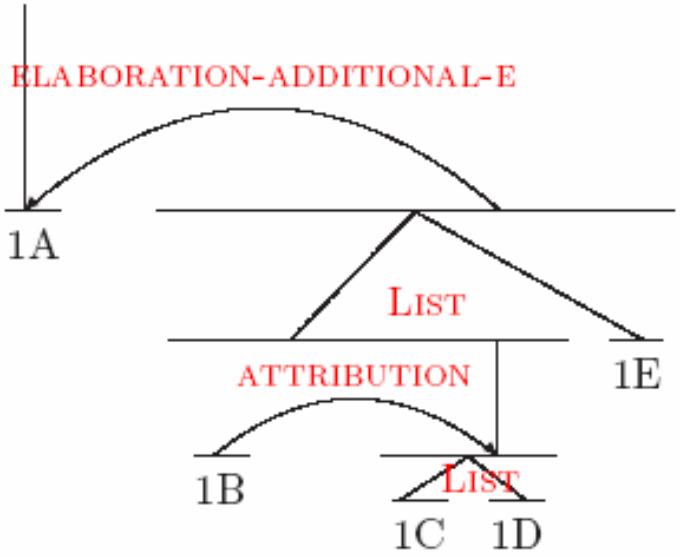
- Introduction to Rhetorical Structure Theory
- Research question
- Data and method
- Results
- Discussion and conclusion
- Questions

Rhetorical Structure Theory

- Mann and Thompson (1988)
- Rhetorical relations between text spans
- Nucleus and Satellite
- Elementary Discourse Units (EDUs)

- RST Treebank (Carlson et al. 2003)
- 385 Wall Street Journal texts

Rhetorical Structure Theory



[Sun Microsystems Inc., a computer maker, announced the effectiveness of its registration statement for \$125 million of 6 3/8% convertible subordinated debentures due Oct. 15, 1999.]^{1A} [The company said]^{1B} [the debentures are being issued at an issue price of \$849 for each \$1,000 principal amount]^{1C} [and are convertible at any time prior to maturity at a conversion price of \$25 a share.]^{1D} [The debentures are available through Goldman, Sachs & Co.]^{1E} (wsj$_{0650}$)

Automatic RST annotation

- Existing systems:
  - Rhetorical Structure Theory Analyser (RASTA, Corston-Oliver 1998)
  - Marcu (1999) and Marcu (2000)
  - Sentence-level PArsing of DiscoursE (SPADE, Soricut & Marcu 2003)
  - Discourse Analysing System (DAS, LeThanh 2004)
  - Timmerman (2007)

Research question

Can we identify features that can be used to predict the presence of rhetorical (RST) relations between (Multi-)Sentential Discourse Units within paragraphs in English?

(Multi-)Sentential Discourse Unit

A (Multi-)Sentential Discourse Unit or (M-)SDU is a text span with a length of at least one sentence and at most one paragraph, forming a discourse unit in a text.

Research question

Can we identify features that can be used to predict the presence of rhetorical (RST) relations between (Multi-)Sentential Discourse Units within paragraphs in English?

Methods

- Finding potentially relevant features
- Extracting them automatically
- Simplifying discourse analysis to a machine learning task
- Interpreting machine learning models

Potentially relevant features

- Literature on existing systems
- Data:
  over 200 relations from 30 different texts
  (RST Treebank)

Potentially relevant features

type	*feature*	*C-O 98*	*M 99*	*M 00*	*LT 04*	*T 07*	*data*
Surface	words		x				
	POS tags		x				
	(M-)SDU length						
Syntactic	syntactic similarity						x
	tense, aspect and polarity	x					
Reference	anaphora resolution	x			x		
	personal pronouns					x	x
	definite articles						x
	demonstrative pronouns						x
	reference words (e.g. further)						x
	(wh-)determiners (e.g. which)						x
	NP simplification						x

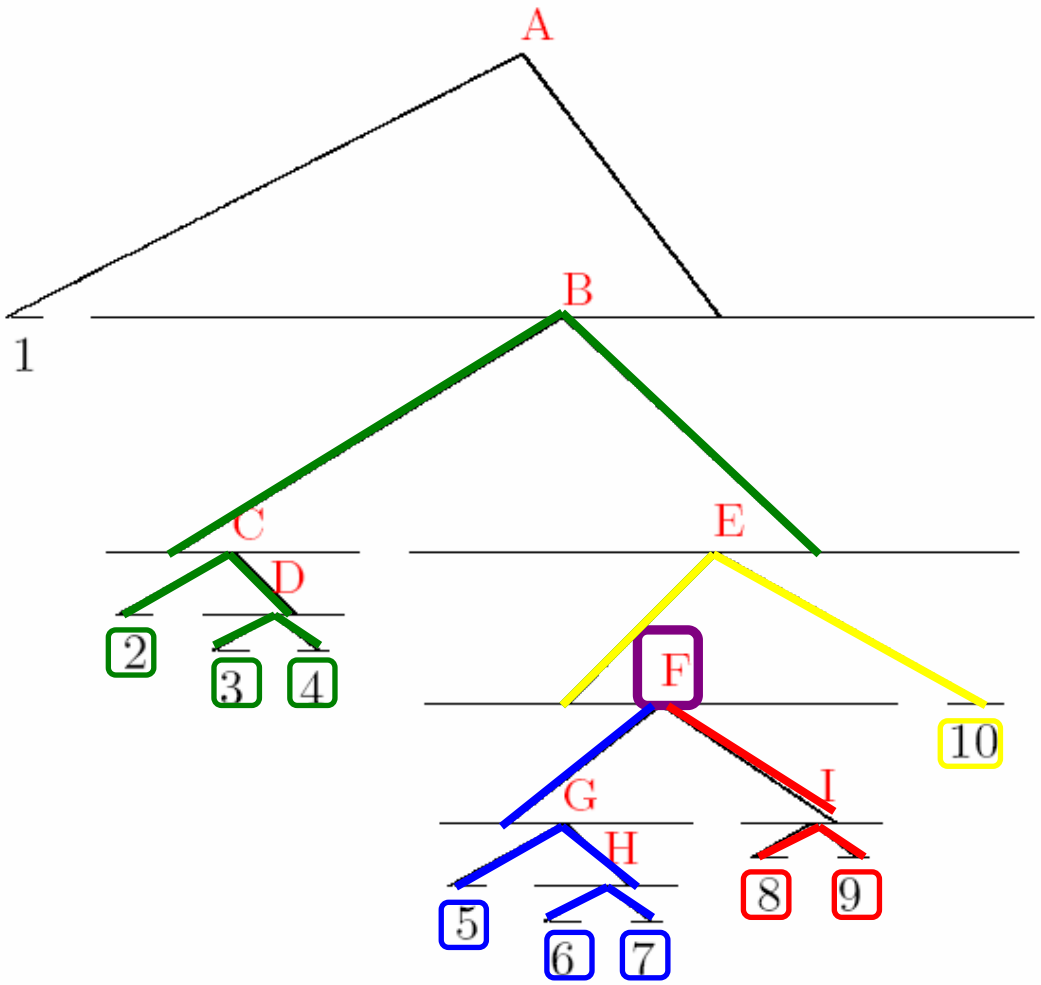
Potentially relevant features

type	*feature*	*C-O 98*	*M 99*	*M 00*	*LT 04*	*T 07*	*data*
Lexical	cue phrases	x	x	x	x	x	x
	NP and VP cues				x		
	word overlap		x	x		x	x
	word similarity		x		x	x	x
	time references				x		x
Discourse	position in the text						
	continuous punctuation						x
	internal discourse structure						
Meta-info	newspaper style						x
	world knowledge						x

Extracting features automatically

- Treebanks and Thesauri
  - RST Treebank (Carlson et al. 2003)
  - Penn Treebank (Marcus et al. 1993)
  - WordNet-2.1 (Fellbaum 1998)
  - CELEX (Baayen et al. 1993)
  - Dependency-Based Thesaurus (Lin 1998)
- Tools
  - WordNet-Similarity-1.04 (Pedersen et al. 2004)
  - GuiTAR (Poesio & Alexandrov-Kabadjov 2004)

Simplifying discourse analysis



For each relation:

1. Find possibly related text spans on the left
2. Same for right
3. For each triple deceide whether relation is on the left or right

Data

- 2136 triples extracted from RST Treebank
  - *right* 1196
  - *left* 940

- 1836 features
  - best surface features according to Relief: 1000
  - syntactic features: 20
  - reference features: 84
  - lexical features: 718
  - discourse features: 14

Machine learning algorithms

- Orange (Demsar et al. 2004)
  - Naive Bayes
  - K-Nearest Neighbours (kNN)
  - Support Vector Machines (SVM)
  - Decision Trees

- Maximum Entropy (Zhang 2004)

Testing

- ten-fold cross-validation
- Manual separation into partitions
- Accuracy =

$$\frac{\text{total of correctly classified cases}}{\text{number of cases (2136)}}$$

- Baseline is 56.0% (choosing *right*)

Results

Algorithm	*accuracy*
Naive Bayes	60.0%
kNN	51.1%
SVM	56.9%
Decision Trees	53.1%
Maximum Entropy	60.9%

Baseline = 56.0%

Finding relevant features

- Models of machine learning algorithms
  - Naive Bayes
  - Maximum Entropy

- Feature selection algorithms
  - Relief (Kira & Rendell 1992; Kononenko 1994)
  - Cluster Separation Score (CSS, van Halteren)

Useful features

- Present in more than one partition
- 50 best features following ranking by each of the four algorithms

Surface features

- Words
  - *a* middle
  - *as* middle
  - *farmer* right
  - *it* * right
  - *little* middle
  - *the* ** right
  - *to* middle

- POS tags
  - personal pronoun middle, right
  - proper noun, sg.** right
  - verb, 3rd p. sg. pr.*** left

* *left* and *middle* are in the top 100
** *middle* is in the top 100
*** *right* is in the top 100

Syntactic features

- Syntactic similarity
  - syntactic similarity LEFT
- Tense, aspect and polarity
  - infinitive * left, middle
  - past tense * left
  - present tense left, right
  - gerund ** middle
  - modal *** right

* *right* is in the top 100
** *left* is in the top 100
*** *left* and right are in the top 100

Reference features (1)

- Anaphora
  - Anaphora LEFT and RIGHT
- Personal pronouns
  - number of pers. pronouns LEFT and RIGHT
  - pers. pronoun in first clause LEFT and RIGHT
- Definite articles
  - number of def. articles RIGHT
  - def. article in first clause * RIGHT
- Demonstrative pronouns
  - number of dem. pronouns * RIGHT

* *LEFT* is in the top 100

Reference features (2)

- Reference words
  - *further* LEFT
  - *less* LEFT and RIGHT
  - *other* LEFT and RIGHT
  - *added* * RIGHT
  - *more* RIGHT
- NP simplification
  - missing modifier LEFT and RIGHT

* *LEFT* is in the top 100

Lexical features

- Word overlap
  - token overlap * RIGHT
  - lemma overlap RIGHT
  - stem overlap * RIGHT
- Word similarity
  - Lin's dep. thesaurus LEFT and RIGHT
- Time references
  - time references right

* LEFT is in the top 100

Discourse features

- Position in the text
  - text (paragraph nr) — left, middle, right
  - paragraph (sent. nr) * — right, middle
- Continuous punctuation
  - quotation marks — RIGHT
- Internal discourse structure
  - discourse structure * — RIGHT

* left is in the top 100
** LEFT is in the top 100

Conclusion

- We have found relevant features
  - Disappointing results (due to artificial task?)
  - Text genre specific (news paper)

- Suggestions for future research
  - Apply the features to the real task
  - Employ larger data sets in different text genres
  - Consider feature interaction

Questions?

Top 10

1.	Pers. pronoun 1st cl.	RIGHT	reference
2.	Def. article 1st clause	RIGHT	reference
3.	Quotation marks	RIGHT	discourse
4.	Past tense	left	syntactic
5.	Token overlap	RIGHT	lexical
6.	Pers. pronoun	right	surface
7.	Time reference	right	lexical
8.	Missing modifier	RIGHT	reference
9.	Present tense	left	syntactic
10.	Lemma overlap	RIGHT	lexical