Constrained decoding for text-level discourse parsing

Philippe Muller[1] *Stergos Afantenos*[1] *Pascal Denis*[2] *Nicholas Asher*[3]

(1) IRIT, Université de Toulouse, France
(2) Mostrare, INRIA, France
(3) IRIT, CNRS, France
{stergos.afantenos,muller,asher}@irit.fr, pascal.denis@inria.fr

Coling 2012, Mumbai
December 2012

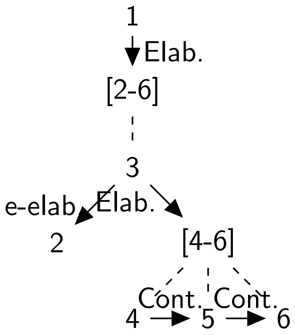
Big picture

- Discourse analysis = discourse units + relations between units
- Discourse parsing = finding relations, given units
- relations = unit pair + label
- label = "rhetorical" function:
  explanation, elaboration, contrast, continuation, ...
- why ? thematic structure + implicit semantic pieces of information

Example

[Principes de la sélection naturelle.]_1 [La théorie de la sélection naturelle [telle qu'elle a été initialement décrite par Charles Darwin,]_2 repose sur trois principes:]_3 [1. le principe de variation]_4 [2. le principe d'adaptation]_5 [3. le principe d'hérédité]_6

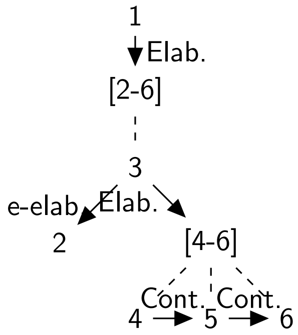
[Principles of natural selection.]_1 [The theory of natural selection, [as it was initially described by Charles Darwin]_2, lies upon three principles:]_3 [1. the principle of variation]_4 [2. the principle of adaptation]_5 [3. the principle of heredity]_6



Example

[Principes de la sélection naturelle.]_1 [La théorie de la sélection naturelle [telle qu'elle a été initialement décrite par Charles Darwin,]_2 repose sur trois principes:]_3 [1. le principe de variation]_4 [2. le principe d'adaptation]_5 [3. le principe d'hérédité]_6

[Principles of natural selection.]_1 [The theory of natural selection, [as it was initially described by Charles Darwin]_2, lies upon three principles:]_3 [1. the principle of variation]_4 [2. the principle of adaptation]_5 [3. the principle of heredity]_6

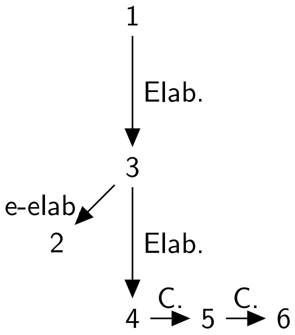


some complex structure

Example

[Principes de la sélection naturelle.]_1 [La théorie de la sélection naturelle [telle qu'elle a été initialement décrite par Charles Darwin,]_2 repose sur trois principes:]_3 [1. le principe de variation]_4 [2. le principe d'adaptation]_5 [3. le principe d'hérédité]_6

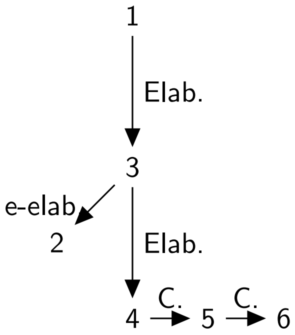
[Principles of natural selection.]_1 [The theory of natural selection, [as it was initially described by Charles Darwin]_2, lies upon three principles:]_3 [1. the principle of variation]_4 [2. the principle of adaptation]_5 [3. the principle of heredity]_6



Example

[Principes de la sélection naturelle.]_1 [La théorie de la sélection naturelle [telle qu'elle a été initialement décrite par Charles Darwin,]_2 repose sur trois principes:]_3 [1. le principe de variation]_4 [2. le principe d'adaptation]_5 [3. le principe d'hérédité]_6

[Principles of natural selection.]_1 [The theory of natural selection, [as it was initially described by Charles Darwin]_2, lies upon three principles:]_3 [1. the principle of variation]_4 [2. the principle of adaptation]_5 [3. the principle of heredity]_6



or a simple labelled graph

Discourse parsing

- given the units, find which ones are related
  ("attachment" problem)
- optionally, group them in complex units
- label relations with their rhetorical function, the author's "intention"
  ("labelling" problem)

Main issues:

- data sparsity
- interdependence between attachments → global constraints on well-formedness (not settled theoretically)
- interdependence between attachment and labelling

Frameworks and Data

- theories in competition with different structural assumptions:
  - Rhetorical Structure Theory: trees, contiguous complex segments
  - Segmented Discourse Representation Theory: multi-graph, complex units, some constraints on attachment
  - Wolf & Gibson: multi-graph, complex units, no constraints on attachment
- Corpora:
  - RST treebanks in English ($>$1), Spanish
  - SDRT (Discor, English) or SDRT-like (Annodis, French)
  - Wolf & Gibson (English)

Frameworks and Data

- theories in competition with different structural assumptions:
  - Rhetorical Structure Theory: trees, contiguous complex segments
  - Segmented Discourse Representation Theory: multi-graph, complex units, some constraints on attachment
  - Wolf & Gibson: multi-graph, complex units, no constraints on attachment
- Corpora:
  - RST treebanks in English ($>$1), Spanish
  - SDRT (Discor, English) or SDRT-like (Annodis, French)
  - Wolf & Gibson (English)

$\rightarrow$ we go towards a common (partial) representation, simple dependency graphs with general decoding strategy

Frameworks and Data

- theories in competition with different structural assumptions:
  - Rhetorical Structure Theory: trees, contiguous complex segments
  - Segmented Discourse Representation Theory: multi-graph, complex units, some constraints on attachment
  - Wolf & Gibson: multi-graph, complex units, no constraints on attachment
- Corpora:
  - RST treebanks in English ($>$1), Spanish
  - SDRT (Discor, English) or SDRT-like (Annodis, French)
  - Wolf & Gibson (English)

$\rightarrow$ we go towards a common (partial) representation, simple dependency graphs with general decoding strategy
then: adjust your constraints for well-formed structures, optimize predictions wrt these constraints

Discourse parsing

Past approaches:

- local models learnt
- greedy heuristics-based decoding and/or corpus specific features
- tree-structure
- english corpora: RST treebanks, Verbmobil

Our approach:

- elementary units only
- dependency graph
- local model(s) but decoding with global constraints on the structure, and global optimization of the result
- tested on French Annodis Corpus

Decoding strategies

Depending on the structure aimed at

- greedy local attachments (Duverlé & Prendinger)
- transformation-based parsing to yield trees (di Eugenio, Sagae) cf shift-reduce in syntax
- ours:
  - maximal spanning tree, cf dependency parsing in syntax = unconstrained tree
  - global optimization of the structure probability with A* and custom constraints
- strong baseline in all corpora: attachment of each unit to the previous one

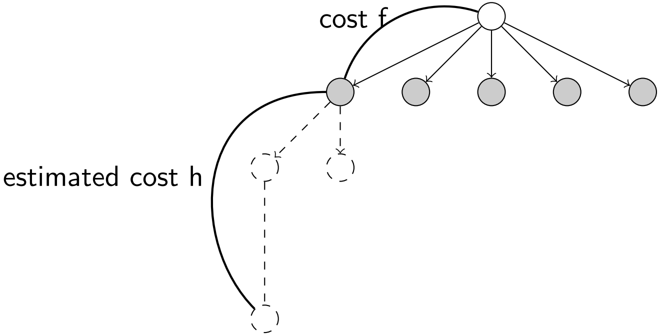
A^* search I

- shortest path search through the state-space of possible results = possible discourse structures, built incrementally
- at every decision point, order all continuations based on a "cost", summing
  - cost of the partial solution already built
  - an estimated cost of what remains to be done

  keep every option open (contra beam search) and start with the lowest cost
- "cost" related to probabilities of structures, must be additive, ≥ 0 and lower is better: $-log(p)$

A* search II

gray = decision points



value of considered node = f+h

A^* search for discourse parsing

state-space exploration is incremental; the following should be defined:

- the start state
- allowed states from a given state
- an estimation function for the cost

A* search for discourse parsing

state-space exploration is incremental; the following should be defined:

- the start state e.g. first elementary discourse unit
- allowed states from a given state

- an estimation function for the cost

A^* search for discourse parsing

state-space exploration is incremental; the following should be defined:

- the start state e.g. first elementary discourse unit
- allowed states from a given state
  e.g. link a DU to exactly one already introduced DU ($\rightarrow$ tree)
- an estimation function for the cost

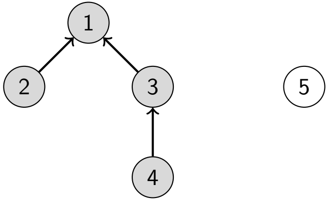
A* search for discourse parsing

state-space exploration is incremental; the following should be defined:

- the start state e.g. first elementary discourse unit
- allowed states from a given state
  e.g. link a DU to exactly one already introduced DU ($\rightarrow$ tree)
- an estimation function for the cost
  e.g. average of linking cost for every remaining DU

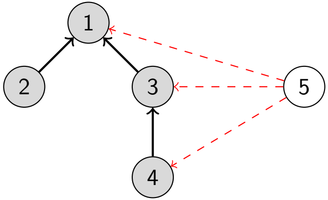
Constraints on structures

other constructions will yield different kinds of structures:



Constraints on structures

other constructions will yield different kinds of structures:
e.g. restricting linking sites to most recent nodes "higher up" on the tree, a.k.a. "right frontier constraint" [Polanyi, 1988]



relation name	#	%	relation name	#	%
alternation	18	0.5	explanation	130	3.9
attribution	75	2.2	flashback	27	0.8
background	155	4.6	frame	211	6.3
comment	78	2.3	goal	95	2.8
continuation	681	20.3	narration	349	10.4
contrast	144	4.3	parralel	59	1.8
E-elab	527	15.7	result	163	4.9
elaboration	625	18.6	temploc	18	0.5
total # relations		3355	total # EDUs		3188
total # CDUs		1395	total # texts		86

Relations can be grouped into 4 main classes:

- structural
- sequence
- expansion
- temporal

Experiments
Annodis Corpus

relation name	#	%	relation name	#	%
alternation	18	0.5	explanation	130	3.9
attribution	75	2.2	flashback	27	0.8
background	155	4.6	frame	211	6.3
comment	78	2.3	goal	95	2.8
continuation	681	20.3	narration	349	10.4
contrast	144	4.3	parralel	59	1.8
E-elab	527	15.7	result	163	4.9
elaboration	625	18.6	temploc	18	0.5
total # relations		3355	total # EDUs		3188
total # CDUs		1395	total # texts		86

Relations can be grouped into 4 main classes:

- structural
- sequence
- expansion
- temporal

relation name	#	%	relation name	#	%
alternation	18	0.5	explanation	130	3.9
attribution	75	2.2	flashback	27	0.8
background	155	4.6	frame	211	6.3
comment	78	2.3	goal	95	2.8
continuation	681	20.3	narration	349	10.4
contrast	144	4.3	parralel	59	1.8
E-elab	527	15.7	result	163	4.9
elaboration	625	18.6	temploc	18	0.5
total # relations		3355	total # EDUs		3188
total # CDUs		1395	total # texts		86

Relations can be grouped into 4 main classes:

- structural
- sequence
- expansion
- temporal

relation name	#	%	relation name	#	%
alternation	18	0.5	explanation	130	3.9
attribution	75	2.2	flashback	27	0.8
background	155	4.6	frame	211	6.3
comment	78	2.3	goal	95	2.8
continuation	681	20.3	narration	349	10.4
contrast	144	4.3	parralel	59	1.8
E-elab	527	15.7	result	163	4.9
elaboration	625	18.6	temploc	18	0.5
total # relations		3355	total # EDUs		3188
total # CDUs		1395	total # texts		86

Relations can be grouped into 4 main classes:

- structural
- sequence
- expansion
- temporal

- Our discourse parsing is based on two locally-trained classifiers:
  - one predicts the attachment site of each DU
  - the other predicts discourse relation for attached pairs of DUs
- In both cases, we trained two different types of probabilistic model:
  - Naive Bayes
  - Maximum Entropy
- The choice of probabilistic models is guided by the way we combine the two models during decoding
- Models were trained on 10-fold cross validation on the document level

- Features shared by the two classifiers
  - EDU_i and EDU_j in the same sentence or paragraph
  - $\mathrm{EDU}_{i/j}$ is the first EDU in the paragraph
  - Number of tokens in an $\mathrm{EDU}_{i/j}$
  - Number of intervening EDUs between EDU_i and EDU_j
  - Whether the EDU_i is embedded in EDU_j and conversely
- Attachment features
  - Presence of a particular discourse marker
  - EDU_j is embedded in an EDU other than EDU_i
  - $\mathrm{EDU}_{i/j}$ is an apposition or relative clause embedded in its main clause

- Relation labeling features
  - Presence of a verb in $\mathrm{EDU}_{i/j}$
  - Which discourse relations are triggered from all discourse markers in $\mathrm{EDU}_{i/j}$
  - Syntactic category of the head token of $\mathrm{EDU}_{i/j}$
  - Presence of a negation, tense agreement between head verbs of both EDU_i and EDU_j
  - features inspired from coreference resolution (based on pronouns and NPs)

attachment either unconstrained (full) or limited to units in a 5-unit window

	MaxEnt	NB
w5	**67.4**	61.1
full	63.5	51.3

The difference between Maxent and Naive Bayes is significant at $p < 0.01$, using McNemar's test. The upper limit recall for the latter task in w5 configuration is 92%.

	MaxEnt	NB	Majority
w5 (18 rels)	**44.8**	34.7	19.1
full (18 rels)	43.3	32.9	19.7
w5 (4 rels)	**65.5**	62.1	51.2
full (4 rels)	63.6	60.1	50.1

Results 1: attachment of DUs

Training model	Naive Bayes			Maxent		
Decoding method	greedy	MST	A*	greedy	MST	A*
attachment alone (w5)	61.2	65.7	**66.2**	62.1	65.7	65.7
attachment alone	58.5	62.0	62.1	62.2	*65.7*	*65.7*
joint/unlabelled (w5)	59.7	61.7	64.8	62.2	65.1	*65.3*
joint/unlabelled	57.9	57.0	59.6	62.3	65.1	*65.4*

- A* and MST decoding similar, but differ from other methods.
- Confidence intervals at 95% are all about $\pm$ 0.9-1.2% wrt to given scores.

Results 2: labelled graphs

Training model		Naive Bayes			Maxent			
Decoding method		greedy	MST	A*	greedy	last	MST	A*
joint(w5)	4 rels	38.9	29.3	41.7	42.2	42.2	31.6	*44.1*
joint	4 rels	38.7	26.7	39.6	44.6	44.5	30.0	**46.8**
pipe-line(w5)	4 rels	39.5	42.1	42.5	42.1	42.2	*44.3*	*44.3*
pipe-line	4 rels	38.7	40.8	40.8	44.5	44.5	**46.8**	**46.8**
joint(w5)	18 rels	22.0	8.2	23.7	28.7	28.6	4.8	*30.1*
joint	18 rels	23.4	4.1	24.0	34.2	34.1	5.4	**36.1**
pipe-line(w5)	18 rels	22.5	24.0	24.5	28.7	28.6	*30.2*	*30.2*
pipe-line	18 rels	23.9	24.7	24.8	34.0	34.1	**36.1**	**36.1**

- 'last' baseline uses a maxent model for prediction of relations.
- Confidence intervals at 95% are all about ± 2% wrt to given scores.
- Best joint and pipe-lined scores are not significantly different from each other.

Beyond

– data:
translate RST treebanks into dependency graphs to use bigger corpora

– methods

- learning under same constraints as in decoding
- ranking n-best output (given almost for free by A^*)

Polanyi, L. (1988).
A formal model of the structure of discourse.
Journal of Pragmatics, 12.