Exploiting naive vs expert discourse annotations: an experiment using lexical cohesion to predict Elaboration / Entity-Elaboration confusions

Clémentine Adam, Marianne Vergez-Couret

CLLE - Université de Toulouse, France

July 12th 2012

Introduction

- Field: Corpus annotation at the discourse level
- The ANNODIS corpus:
  - top-down approach: annotation of macro structures (enumerative structures)
  - bottom-up approach: construction of the structure of discourse *via* discourse relations between elementary discourse units (EDU)
- Other corpus annotated with discourse relations: RST TreeBank (Carlston 2001), Penn Discourse TreeBank (Prasad 2007), Discor Corpus (Reese 2007)

Introduction

- ANNODIS is the first corpus annotated with discourse relations for french
- It provides two levels of annotations:
  - a pre-annotation done by naive annotators (naive annotation)
  - a revised annotation done by expert annotators (expert annotation)
- Objective of this study: using the two levels, we want to predict confusions between two relations, Elaboration and Entity-Elaboration (E-Elaboration)

Outline

Outline

The ANNODIS Corpus

- Composition: Wikipedia articles and newspaper articles
- Set of discourse relations, adapted from the SDRT model (Asher 2003) and inspired by other discourse models: RST framework (Mann 1987), Linguistic Discourse Model (Polanyi 1988), graphbank model (Wolf 2005) : Alternation, Attribution, Background, Comment, Continuation, Contrast, Elaboration, Entity-Elaboration, Explanation, Flashback, Frame, Goal, Narration, Parallel, Result, Temporal Location

The ANNODIS Corpus

Three steps of annotation:

- Preliminary annotation: 50 texts; 2 annotators (postgraduate students)
  $\rightarrow$ creation of the annotation manual
- Naive annotation: 86 texts; 3 annotators (other postgraduate students)
  Kappa: 0.4 (week to moderate inter-annotator agreement)
- Expert annotation: 42 texts... still in progress

Outline

Elaboration and Entity-elaboration

- Elaborations + E-elaborations: 50% of the naive annotated relations; 35% of the expert annotated relations

		Naive		
		Elab	E-Elab	Total
Expert	Elab	302	70	372
	E-Elab	158	216	374
	Total	460	286	746
Expert	Continuation	70	32	
	Background	32	18	
	Other	150	59	

Elaboration and Entity-elaboration

- The Elaboration relation relates two propositions only if the second proposition describes a sub-state or sub-event of the state or event described in the first proposition. Elaboration also includes exemplification, reformulation and paraphrase cases.
- The E-Elaboration relation relates two segments for which the second one specifies a property of one of the involved entities in the first segment. This property can be important (e.g. identificatory) or marginal.

Elaboration and Entity-elaboration

[La Lausitz, [une région pauvre de l'est de l'Allemagne,]$_1$ [réputée pour ses mines de charbon à ciel ouvert,]$_2$ a été le théâtre d'une première mondiale, mardi 9 septembre.]$_3$ [Le groupe suédois Vattenfall a inauguré, dans la petite ville de Spremberg, une centrale électrique à charbon expérimentale]$_4$ [qui met en œuvre toute la chaîne des techniques de captage et de stockage du carbone]$_5$

[Lausitz, [a poor region in east Germany,]$_1$ [famous for its open air coal mines,]$_2$ was the scene of a world first, on Tuesday September 9th.]$_3$ [The swedish group Vattenfall inaugurated, in the small town of Spremberg, an experimental coal power plant]$_4$ [involving the complete carbon capture and storage chain.]$_5$

E-Elaboration (3,[1-2])
Elaboration (3,4)
E-Elaboration (4,5)

Elaboration and Entity-elaboration

- No prototypical marker exists, neither for Elaboration nor E-elaboration
- Some possible markers are indicated in the ANNODIS manual: *à savoir, c'est-à-dire, notamment*, etc.
  → But they can mark both relations
- Other possible linguistic features:
  - Prévot (2009): E-elaboration can be realized by relative clauses and appositions (nominal and adjectival appositions, brackets...)
  - Vergez-Couret (2012): French gerund clauses may express several discourse relations including Elaboration but not E-elaboration

  → All these features are ambiguous and seldom appear

Outline

Lexical cohesion of Elaboration and E-elaboration

[Un soir, il faisait un **temps** horrible,]$_{16}$ [les **éclairs** se croisaient,]$_{17}$ [le **tonnerre** grondait,]$_{18}$ [la **pluie** tombait à torrent.]$_{19}$

[One night, the **weather** was horrible,]$_{16}$ [**flashes of lightning** were crossing,]$_{17}$ [**thunder** growled,]$_{18}$ [**rain** fell heavily.]$_{19}$

Elaboration (16,[17-19])

Lexical cohesion of Elaboration and E-elaboration

[Pourquoi a-t-on abattu Paul Mariani, [cinquante-cinq ans]$_4$, [attaché au cabinet de M. François Doubin,]$_5$?]$_6$

[Why was Paul Mariani, [fifty-five]$_4$, [personal assistant to M. François Doubin,]$_5$ gunned down?]$_6$

E-elaboration (6,[4-5])

Lexical cohesion of Elaboration and E-elaboration

In order to evaluate the strength of lexical cohesion between two segments S_a and S_b:

1. The two segments are annotated with part-of-speech and lemma information using the TreeTagger (Schmid 1994)
2. All the lexical proximity links between the two segments are annotated. To detect these links, we use a lexical proximity measure based on the distributional analysis of the french Wikipedia (Bourigault 2002)

Calling N_ℓ the number of links between S_a and S_b, N_a and N_b the numbers of words in the segments S_a and S_b, our score Sc is defined as:

$$Sc = \frac{N_\ell}{\sqrt{N_a \cdot N_b}}$$

Lexical cohesion of Elaboration and E-elaboration

	Elab.	E-elab.
Number of cases	625	527
Average # of proj. links N_ℓ	**5.99**	1.39
Average cohesion score Sc	**0.61**	0.32

$\rightarrow$ Elaboration is much more cohesive than E-Elaboration

Outline

Predicting confusions between Elab. and E-elab.

Att.	Description	Values
N_ℓ	Number of links	$N_\ell \in \mathbb{N}$
Sc	Lexical cohesion score	$Sc \in \mathbb{R}^+$
rel	S_b is a relative clause	boolean
app	S_b is a nom. / adj. apposition	boolean
ger	S_b is a gerund clause	boolean
bra	S_b is in brackets	boolean
emb	S_b is an embedded segment	boolean
w_{Sa}	# of words in S_a	$w_{S1} \in \mathbb{N}$
w_{Sb}	# of words in S_b	$w_{S2} \in \mathbb{N}$
w_{tot}	$w_{Sa} + w_{Sb}$	$w_{tot} \in \mathbb{N}$
s_{Sa}	# of segments in S_a	$s_{S1} \in \mathbb{N}$
s_{Sb}	# of segments in S_b	$s_{S2} \in \mathbb{N}$
s_{tot}	$s_{Sa} + s_{Sb}$	$s_{tot} \in \mathbb{N}$

Predicting confusions between Elab. and E-elab.

Naive *vs* Expert annotations:

	elab	e-elab
elab	302	70
e-elab	158	216

← Naive annot.

↑Expert annot.

Accuracy : 69.4%

Predicting confusions between Elab. and E-elab.

Classification using Weka's (Hall 2009) implementation of Random Forest classifier (Breiman 2001):

	elab	e-elab
elab	306	66
e-elab	115	259

← Naive-aided auto. annot.

↑Expert annot.

Accuracy : 75.7%

Predicting confusions between Elab. and E-elab.

Impact of the different attributes' categories:

Attributes used	Accuracy
Naive annotation	69.4%
Naive + lexical cohesion cues	72.3% (+2.9%)
Naive + linguistic cues	71.7% (+2.3%)
Naive + structural cues	69.7% (+0.3%)
All	75.7% (+6.3%)

Predicting confusions between Elab. and E-elab.

Using our classifier to reduce the experts' workload:

Cost matrix:

0	10
1	0

Results:

	elab	e-elab
elab	57	13
e-elab	70	146
↑Expert	127	159

← automatic annot. (naive annot=e-elab)

↙ second look by expert

↘ accepted annot. (error : 8.2%)

Conclusion

- We focused on two frequent discourse relations, Elaboration and E-elaboration, which are:
  - often interchanged by annotators
  - difficult to detect automatically
- Using the ANNODIS corpus, we show that lexical cohesion is a strong cue to differentiate between them
- We used this cue, among others, in a machine learning experiment allowing to reduce the experts' workload for the revision of the naive annotation

Thank you...

... for your attention !