# Learning Recursive Segments for Discourse Parsing

Stergos D. Afantenos* Pascal Denis† Philippe Muller* †
Laurence Danlos†

*Institut de Recherche en Informatique de Toulouse,
CNRS, Université Toulouse III Paul Sabatier
†Equipe-Projet Alpage INRIA & Université Paris 7

LREC 2010

# Goal

- discourse parsing :
  - segmentation : segment a discourse into Elementary Discourse Units (EDUs)
  - linking : link EDUs with rhetorical relations (cf Rhetorical Structure Theory (RST), Segmented Discourse Representation Theory (SDRT)).
- we focus on the first subtask, within the framework of SDRT.
- EDUs in SDRT, in contrast to other theories—*e.g.* RST—are allowed to be embedded.

# Embedded EDUs

## Example from RST corpus :

[But maintaining the key components of his strategy]$_1$ [– a stable exchange rate and high levels of imports –]$_2$ [will consume enormous amounts of foreign exchange.]$_3$

In RST, units 1 and 3 will later be linked with an *ad hoc* "same-unit" relation.
Here we chose to deal with this problem at the segmentation stage.

## Desired segmentation

[But maintaining the key components of his strategy [– a stable exchange rate and high levels of imports –]$_1$ will consume enormous amounts of foreign exchange.]$_2$

# ANNODIS

- The corpus we used was created within the ANNODIS project, an on going effort to build a discourse graph bank for French texts ; it has the following goals :
  1. testing various theoretical proposals about discourse structure, and
  2. providing a seed corpus for learning discourse structures using Machine Learning.
- It aims at creating 100–150 documents, segmented and annotated with discourse relations.

# Corpus

Our experiments have been performed in 47 documents, which have undergone validation, drawn from the ANNODIS corpus.

| *# Docs* | *# Tokens* | *# EDUs* | *% Embedded EDUs* |
|---|---|---|---|
| 47 | 15156 | 1445 | 10% |

# Experiments

We opted for a token-based classification, classifying each token into four classes :

| | |
|---|---|
| **LEFT** | token starts an Elementary Discourse Unit (EDU) |
| **RIGHT** | token ends an EDU |
| **BOTH** | single-token EDU (*e.g.* titles, some frame adverbials) |
| **NOTHING** | none of the above. |

Machine learning based segmentation systems with no embedded EDU (RST) use a binary classification system (boundary or not), with no problem of balanced bracketing

Our segmentation task is akin to clause boundary identification task (CBI) which uses three classes (start, end, inside), with balanced bracketing problem

# Classifier

We used a **(regularized) Maximum Entropy model** :

$$P(b|t) = \frac{1}{Z(b)} \exp\left(\sum_{i=1}^{m} w_i f_i(t, b)\right)$$

- $b$ : the outcome (boundary type)
- $t$ : the token, encoded as a vector of $m$ indicator features $f_i$
- $w_i$ : the weight for $f_i$, with $w$=weight vector,
- $Z(b)$ : normalization factor over the different class labels,
- The values for the parameters $\hat{w}$ are obtained by maximizing the log-likelihood of the training data $T$ with respect to the model :

$$\hat{w} = \underset{w}{\operatorname{argmax}} \sum_{i}^{T} \log P(b^{(i)}|t^{(i)})$$

# Features Set

Our feature set relies on two main sources of information :

- Lexical Features
- Syntactic features, derived from
  - a chunker (Macaon),
  - a full syntactic parser (Syntex)

# Features Set

Lexical Features :

| Feature | Description |
|---|---|
| Lemma | the token's lemma |
| POS | Part of speech |
| Grammatical category | the main grammatical category of the token : V, N, P, etc. |
| start of a discourse marker | boolean, indicating whether the tokens starts a discourse marker |
| indirect speech report verb | boolean, indicating whether the token belongs to a predefined list of verbs. |
| distance from sentence boundaries | the relative distance from each of the sentence boundaries |
| context 3-grams | the lemma and POS 3-grams before and after the token |

# Features Set

Syntactic Features :

| | |
|---|---|
| dependency path | the dependency path from the word towards the root, limited to distance 3 (Syntex) |
| inbound dependencies | the inbound dependency relations for each token (Syntex) |
| syntactic projections | the number of times that the token is at the start, end or middle of an NP, VP, PP projection (Syntex) |
| chunk start/end | boolean features ; token coincides with a chunk start/end (Macaon) |
| outward chunk tag sequence | the sequence of chunk tags from the innermost to the outermost chunk (Macaon) |
| context n-gramms | all the n-gramms ($1 < n \leq 6$) that include the token and do not exceed the limits of the sentence. The n-grams include Lemmas (Synte POS tags (Macaon) and Chunk tags (Macaon) |

# Training instances selection

The distribution of boundary types is heavily skewed towards *N* (Nothing) with 12. 000 instances against 1.400 for each *L* and *R*, so we used a resampling method :

- Tokens inside chunks are never EDU boundaries $\Longrightarrow$ they were removed from the training set and they were tagged directly as *N* on testing.
- Tokens at the sentence boundaries are always *L* and *R* $\Longrightarrow$ we kept them for training but they were tagged directly as *L* and *R* on testing.

After these modifications, we had 9.200 *N* and 1.400 for each *L* and *R*.

# Enforcing Coherence

Token-based local classification does not guarantee the well-formedness of EDUs.

We performed post-processing to balance the bracketing with a two-pass (left-to-right and right-to-left) heuristic on each sentence in order to spot misclassifications.

- For the left-to-right pass, we counted the unbalanced opening brackets, and we correctly classified them.

  [X X] X X X X X] → [X X] [X X X X X]

- For the right-to-left pass, we counted the unbalanced closing brackets, and we correctly classified them.

  [X X X X [X X X] → [X X X X] [X X X]

# Results

Evaluation after 10-fold cross-validation :

**Without post-processing**

| Class | R | P | F |
|---|---|---|---|
| Left | 0.845 | 0.891 | 0.868 |
| Right | 0.881 | 0.925 | 0.902 |
| Both | 0.684 | 0.812 | 0.742 |
| EDUs | 0.427 | 0.880 | 0.575 |

**With post-processing**

| Class | R | P | F |
|---|---|---|---|
| Left | 0.876 | 0.880 | 0.878 |
| Right | 0.885 | 0.889 | 0.888 |
| Both | 0.684 | 1.0 | 0.812 |
| EDUs | 0.719 | 0.748 | 0.733 |

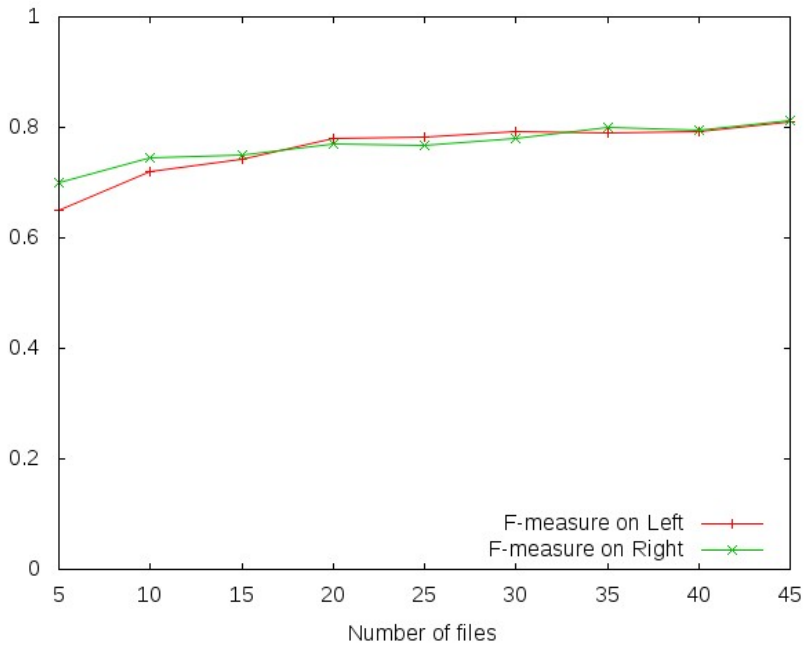
# Comparison with related work

- Non-embedded case on RST (Sagae 2009), EDU F-score= 0.86, which is quite better (but the same-unit relation is not computed yet)
- Our results are close to what can be expected if the problem is seen as a special case of Clause Boundary Identification (CBI) (Marquez 2003) = 0.84
- The number of documents we have been working on (47 documents) is limited.

# Learning curve

- We calculated the learning curve from the 47 documents, in order to see how our approach will benefit from more documents
- We started with 5 documents and we were incrementally adding 5 more documents.
- At each step we performed a 10-fold cross-validation.

# Learning curve




- The curve grows regularly for both classes between sets 5 to 30
- It plateaus between sets 30 and 40
- It grows again during the last set of documents
- It seems that the addition of more documents will only slightly increase our performance.

# Future work

- More global learning models and/or inference procedure (e.g., with local optimization techniques)
- Joint learning of chunking and EDU segmentation
- Assess speed-up during human annotation
- Open question : is it better to learn a “same unit” relation during the segmentation task or the linking task ?