Annotating and learning models of discourse structure

Nicholas Asher
Joint work with Antoine Venant, Philippe Muller, Pascal Denis, Stergos Afantenos, Farah Benamara

IRIT UMR 5505 équipe Mélodi

30 mars 2015

discourse structure

a discourse or dialogue as a structured object (composed recursively out of elementary discourse units EDUs and coherence relations) with semantic effects

- coherence relations—John fired the man who embezzled the funds (embezzle $<$ fire)
- structure— Mary gets stuff done because she's organized. John doesn't $\emptyset_1$ because he doesn't $\emptyset_2$.
- besides ellipsis, structure has semantic/pragmatic effects for pronominal anaphora, scalar implicature, treatment of presuppositions...

Constraints

- In a coherent discourse, every EDU (except maybe one) should be linked to another via some coherence relation.
- Can substructures of the discourse serve as arguments via coherence relations (CDUs) ?
- attachment principles : long distance dependency, right frontier, dialogue turn constraints...
- how many incoming edges to an EDU can there be ? How many rhetorical functions can an EDU have ?

Example

We bought the apartment, <u>but</u> <u>then</u> we rented it out.

Using different corpora with different annotation schemes

Theories of discourse structure with different underlying structures (trees, hierarchical graphs. . .)

- SDRT (multiple incoming arcs, unique root, CDUs, right frontier)
- Graph Bank (multiple arcs)
- discourse dependency trees [Muller&al. 2012, [Li&al. 2014]] (unique incoming arc, long distance)
- RST (adjacency, unique incoming arc, CDUs, right frontier, long distance ?)
- D-LTAG

Questions about this plurality

Q1 : How do the structures compare ?
Venant et al., Expressivity and comparison of models of discourse structures, *Sigdial 2013*.

- Several discourse Annotated Corpora :

RST	:	RST Treebank
SDRT	:	Annodis, Discor, STAC.
others	:	Graphbank(≈less constrained SDRSs), PDTB (≈dependency trees) ,

Q2 : can we translate from one corpus to another, thus extending the range of data available for performing automated tasks ?

- Addressing Q1 is an important step toward Q2.

Learning discourse structures

Different styles of annotated corpora, different tasks :

- learning attachments,
- learning features like nuclearity for argument scope,
- learning discourse relations given attachment points
- learning full structures
- In learning or predicting discourse structures with different annotation schemes are we learning the same thing ?
- transferring from one corpus to another avoids overfitting of features.

Towards Comparison

- Two structures S_1, S_2 express the same predications if there is a bijection μ from relations in S1 to relations in S2 such that the "actual" semantic scope given to relation R is the same as to $\mu(R)$.
- Two possibilities for reading semantic scopes from representations :
  1. what the representation directly expresses—e.g., SDRT
  2. Computed from the structure—e.g., RST with the **Nuclearity Principle** (Marcu1996)

Illustration–SDRT

Recursive Structure :

a SDRS is constitued by discourse units(DUs) linked by discourse relations where a discourse unit is either an elementary unit or a **complex discourse unit (containing a substructure of lower level units.**)
CDUs may not partially overlap.

Graph representation

A SDRS can be seen as a directed acyclic graph where each DU is a vertex and

Directed labelled edges for rhetorical relations.

Directed unlabelled edges link CDUs to their content.

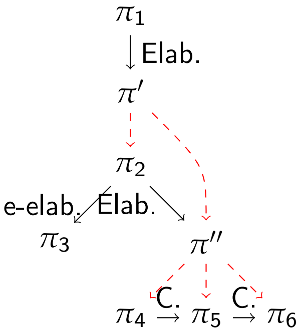
two kind of relations :

coordinating e.g. Result, Narration. Drawn horizontally

subordinating e.g. Elaboration, explanation. Drawn vertically

Illustration–SDRT 2/2

[Principes de la sélection naturelle.]π_1 [La théorie de la sélection naturelle [telle qu'elle a été initialement décrite par Charles Darwin,]π_3 repose sur trois principes :]π_2 [1. le principe de variation]π_4 [2. le principe d'adaptation]π_5 [3. le principe d'hérédité]π_6



Semantic Scopes

The computation of the semantic scopes is transparent :

- $[\![R(\alpha,\beta)]\!] = [\![R]\!]([\![\alpha]\!],[\![\beta]\!])$.
- Scopes here can be described as : $C.(\pi_4,\pi_5) \wedge C.(\pi_5,\pi_6) \wedge Elab.(\pi_3,\pi_2) \wedge e - Elab.(\pi_3,\pi_2) \wedge Elab(\pi_3,[\pi_4,\pi_5,\pi_6]) \wedge Elab(\pi_1,[\pi_3,\pi_4,\pi_5,\pi_6])$.

Illustration—Dependency Trees

- calculate a head for each CDU in an SDRS
- all relations attaching to a CDU attach to its head.

$$
\begin{tikzcd}
 & \pi_1 \arrow[d, "\text{Elab.}"] & & & \\
 & \pi_2 \arrow[dl, "\text{e-elab.}"'] \arrow[dr, "\text{Elab.}"] & & & \\
\pi_3 & & \pi_4 \arrow[r, "\text{C.}"] & \pi_5 \arrow[r, "\text{C.}"] & \pi_6
\end{tikzcd}
$$

Illustration–RST

Two types of relations

- Nucleus-Satellite relations
- Multinuclear Relations

RS-Tree

- An EDU is a RS Tree.
- for a nucleus-statellite relation label R, s_1 and s_2 trees over contiguous spans and $\langle a_1, a_2 \rangle \in \{\langle N, S \rangle; \langle S, N \rangle\}$, $R(t_1\_a_1, t_2\_a_2)$ is an RS Tree.
- for a multinuclear relation label R and $\langle s_1, \ldots, s_n \rangle$ n RS Trees over contiguous spans, $R(s_1\_N, \ldots, s_n\_N)$ is an RS Tree.

Illustration–RST 2/2

Example

[Interprovincial Pipe Line Co. said]π_1 [it will delay a proposed two-step, 830 million dollar [(US$705.6 million)]$\pi_3$ expansion of its system]π_2 [because Canada's output of crude oil is shrinking.]π_4

attribution
n s
π_1 explication
n s
restatement π_4
n s
π_2 π_3

Semantic Scopes

Immediate interpretation :

- $[\![R(t_1, t_2)]\!] = [\![R]\!]([\![t_1]\!], t_2]\!]))$.
- $Restatement(\pi_2, \pi_3) \wedge Explanation([\pi_2, \pi_3], \pi_4) \wedge Attribution(\pi_1, [\pi_2, \pi_3, \pi_4])$.

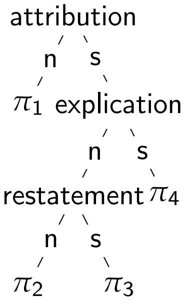
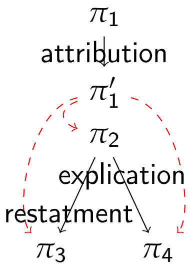
Mixed Nuclearity Principle :

- NS relations only transmit nucleus argument to a parent relation.
- $Restatement(\pi_2, \pi_3) \wedge Explanation(\pi_2, \pi_4) \wedge Attribution(\pi_1, \pi_2)$.

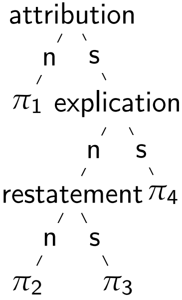
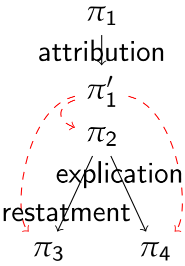
Overview

Same example :

[Interprovincial Pipe Line Co. said]π_1 [it will delay a proposed two-step, 830 million dollar [(US$705.6 million)]$\pi_3$ expansion of its system]π_2 [because Canada's output of crude oil is shrinking.]π_4

RST	SDRT
	

Overview (2/2)

RST	SDRT
	
Mixed NP	immediately
$\mathrm{attr}(\pi_1, \pi_2)$	$\mathrm{attr}(\pi_1, [\pi_2, \pi_3, \pi_4])$
$\mathrm{rest}(\pi_2, \pi_3)$	$\mathrm{rest}(\pi_2, \pi_3)$
$\mathrm{expl}(\pi_2, \pi_4)$	$\mathrm{expl}(\pi_2, \pi_4)$

Intermediate conclusions

- Comparisons require a language expressive enough to express semantic scopes for all discourse theories and interpretative mappings from the different structures into this language.
- The Idea behind an expression like $elab(1, [2, 3]) \wedge elab(2, 3)$: two different instances of an elaboration relation : e_1 and e_2. e_2 scopes over 2 on the left and 3 on the right. e_1 scopes over 1 on the left and both 2 and 3 on the right.
- If RST is interpreted indirectly, the question of the meaning of the initial tree-descendence relation has to be addressed.

Expressing semantic scopes

Semantics scopes are described by mean of a scope structure :

- A scope structure is given by a list of relations instances r_1, r_2, ..., and EDUs intances $\pi_1, \ldots$.
- For each relation name (elaboration, narration, ...), and type (nucleus satellite, multinucleic) a set of relations instances is given to encode information about each relation instance.
- Inclusion in left and right scope respectively, are given by a binary relation $\in_{l/r}$ over relations instances $\times$ Edus.

Using a monadic second order language over this signature, we axiomatize (assuming finite models) RST, SDRT and Depedency trees, as well as encoding and decoding algorithms.

Example

Scope structure

$$elab(1,[2,3]) \wedge elab(2,3) \cong \begin{cases} D : \{r_1, r_2\} \\ ||elab|| = \{r_1, r_2\}, ||expl.|| = \emptyset, \ldots \\ || \in_l || = \{\langle r_1, 1\rangle; \langle r_2, 2\rangle\} \\ || \in_r || = \{\langle r_1, 2\rangle; \langle r_1, 3\rangle; \langle r_2, 3\rangle\} \end{cases}$$

L/R Strong dominance over relation instances :

Let $scopes(r, x)$ stand for $x \in_l r \vee x \in_r r$

$$r \sqsubseteq_{l/r} r' \equiv \forall x scopes(r', x) \rightarrow x \in_{r/l} r$$

Axiomatisation of RST for instance mainly requires specifying that strong dominance yields a tree-like relation.

Weak Dominance

Definition :

- A set elementary constituents X immediately dominates another Y in case $Y \subseteq X$ or there is a subordinating (nucleus-satellite) relations from X to Y. Dominance between constituent is given by transitive closure of this relation.
- r left-weakly dominates r' in case every argument of r' is dominated by the left-arguments of r.

Example :

$$\begin{tikzcd} & \pi_2 \arrow[dl, "\text{rest.}"'] \arrow[dr, "\text{expl.}"] & \\ \pi_3 & & \pi_4 \end{tikzcd}$$

the Restatement instance is left dominated by the explanation one.

nuclearity and dominance

- Interpreting the descendance relation in RST as strong dominance yield the immediate interpretation.
- Mixed NP. fully determine semantic scopes but is not sound.

Our proposal : Relaxed N.P.

- Interpret the descendance relation as weak dominance.
- Unlike strong dominance, there is more than one way of interpreting semantic scopes. Under this view, an RS Tree is an **Underspecified** representation of semantic scopes.
- Choice is left between the combination nucleus-satellite or the nucleus alone, but satellite alone is excluded.

Example

Recall our example :

[Interprovincial Pipe Line Co. said]π_1 [it will delay a proposed two-step, 830 million dollar [(US$705.6 million)]$\pi_3$ expansion of its system]π_2 [because Canada's output of crude oil is shrinking.]π_4

```
        attribution
          /   \
         n     s
        /       \
      π1   explication
               /   \
              n     s
             /       \
     restatement      π4
          /   \
         n     s
        /       \
      π2         π3
```

- The Relaxed N.P. delivers four possible set of interpretations :

attr(π_1, π_2)	attr$(\pi_1, [\pi_2\text{-}\pi_4])$	attr$(\pi_1, [\pi_2, \pi_4])$	attr$(\pi_1, [\pi_2\text{-}\pi_4])$
rest(π_2, π_3)	rest(π_2, π_3)	rest(π_2, π_3)	rest(π_2, π_3)
expl(π_2, π_4)	expl(π_2, π_4)	expl(π_2, π_4)	expl$([\pi_2, \pi_3], \pi_4)$

Example 2/2

These interpretations respectively correspond to :

- The M.NP. interpretation of the RS Tree
- The SDRS for this discourse we saw earlier.
- another sort of DAG
- The immediate interpretation of the RS Tree.

Only the second one is semantically sound, but each of this structure respect the constraints of weak dominance.

back to Expressivity

The following properties hold :

- Immediate interpretation of a RS-Tree can be decoded back to a SDRS which entails that RST+Immediate interpretation is strictly less expressive than SDRT.
- Any interpretation among the set of interpretations of a RS Tree under Relaxed N.P. can be decoded back to a SDRS which yields :
  - An interpretation of a RS Tree as a set of SDRSs.
  - Since the M.N.P interpretation is always a member of the Relaxed N.P. interpretation, the M.N.P interpretation, RST+MNP is strictly less expressive than SDRT.

Intermediate Conclusions

- Both RST and DT provide a kind of underspecified semantic representation of scopes of discourse relations
- RST sometimes specifies several discourse structures for one DT structure
- DT can represent structures that RST cannot (flat structures involving several coordinating relations).
- we now have a way of translating between various discourse structures.
- full equivalence would depend on examining the semantics of different strategies for labeling arcs in the MSO structures.

architecture for parsing

- corpus and feature extraction
- segmentation of text into EDUs
- attachment and labeling of attachment
- local model (probabilities of attachments between all EDUs and probabilities of labels)
- decoders (to find an appropriate structure, the local model with a cutoff value can also serve as a decoder)

Decoding strategies

Depending on the structure aimed at

- greedy local attachments (Duverle & Prendinger)
- transformation-based parsing to yield trees (di Eugenio, Sagae) cf shift-reduce in syntax
- ours (Muller et al. 2012)
    - maximal spanning tree (MST), cf dependency parsing in syntax = unconstrained tree
    - global optimization of the structure probability with A^* and custom constraints like RF
- strong baseline in all corpora : attachment of each unit to the previous one

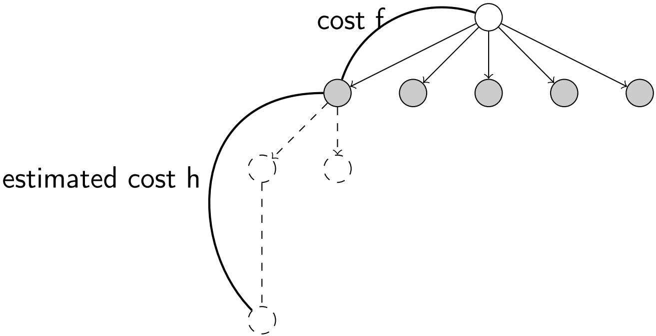
A^* search I

- 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

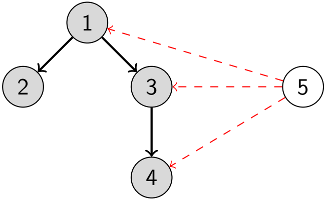
A^* search for discourse parsing

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

- the start state 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 :
e.g. restricting linking sites to most recent nodes "higher up" on the tree, a.k.a. "right frontier constraint" [?]



Experiments

Annodis Corpus

- 18 relations
- grouped into 4 main classes

Experiments

Local classifiers

- 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

Experiments

Feature space

- Features shared by the two classifiers
  - EDU_i and EDU_j in the same sentence or paragraph
  - $\text{EDU}_{i/j}$ is the first EDU in the paragraph
  - Number of tokens in an $\text{EDU}_{i/j}$
  - 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
  - $\text{EDU}_{i/j}$ is an apposition or relative clause embedded in its main clause

Experiments

Feature space (cont'd)

- Relation labeling features
  - Presence of a verb in $\text{EDU}_{i/j}$
  - Which discourse relations are triggered from all discourse markers in $\text{EDU}_{i/j}$
  - Syntactic category of the head token of $\text{EDU}_{i/j}$
  - 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)

Results 1 (ANNODIS) : attachment of DUs

Training model		Naive Bayes			Maxent		0
Decoding method	greedy	MST	A*	greedy	MST	A*	Last
attachment alone (w5)	61.2	65.7	**66.2**	62.1	65.7	65.7	62.4
attachment alone	58.5	62.0	62.1	62.2	65.7	65.7	62.4
joint/unlabelled (w5)	59.7	61.7	64.8	62.2	65.1	65.3	62.4
joint/unlabelled	57.9	57.0	59.6	62.3	65.1	65.4	62.4

- 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 (ANNODIS) : labelled graphs with Maxent

Decoding method		greedy	last	MST	A*
joint(w5)	4 rels	42.2	42.2	31.6	*44.1*
joint	4 rels	44.6	44.5	30.0	**46.8**
pipe-line(w5)	4 rels	42.1	42.2	*44.3*	*44.3*
pipe-line	4 rels	44.5	44.5	**46.8**	**46.8**
joint(w5)	18 rels	28.7	28.6	4.8	*30.1*
joint	18 rels	34.2	34.1	5.4	**36.1**
pipe-line(w5)	18 rels	22.5	28.6	*30.2*	*30.2*
pipe-line	18 rels	34.0	34.1	**36.1**	**36.1**

Analysis

- ANNODIS a very small dataset.
- features are superficial, but rhetorical label detection with maxent is not bad.
- attachment is a difficult problem. CDU detection remains untackled.
- segmentation is important, need good semantic units for better features.
- Our decoders aren't doing much better than the local model with a cutoff. Need for structured prediction ?

Moving to dialogue : Theory

Conversation as a game of message exchange involving a kind of signaling game :

- X plays ϕ
- Y decodes a message in strategic equilibrium (safety, credibility)
- Y decides what signal to send in return
- X decodes a message.
- ...

Empirical work : the corpus of Settlers

A Do you have rock ?

B I've got lots of wheat

[in fact, B has a rock]

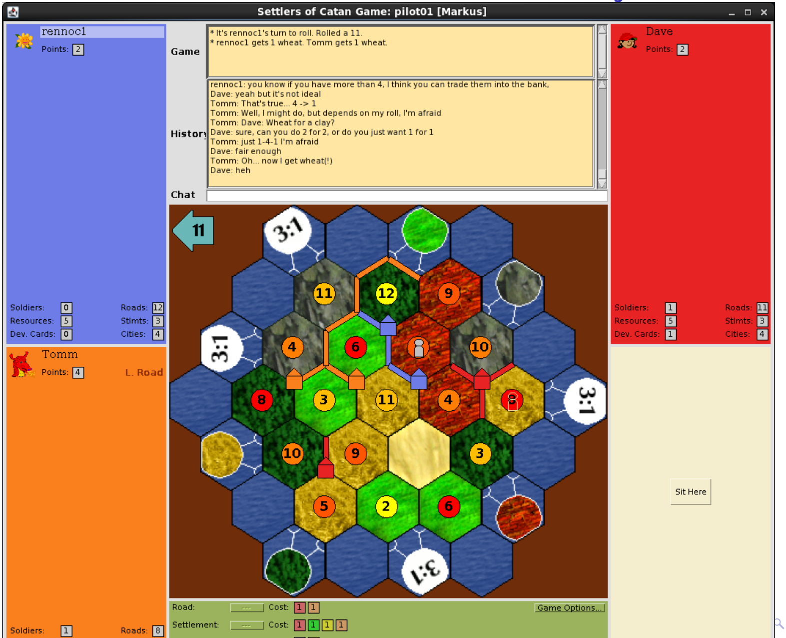
A I'll give you 2 clay for a rock

B How about 2 clay for a wheat ?

A I'll give 1 clay for 3 wheat

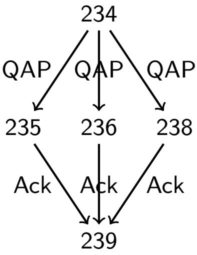
B OK, it's a deal.

Le jeu


Settlers of Catan Game: pilot01 [Markus]
rennocl
Points: 2
Soldiers: 0
Resources: 5
Dev. Cards: 0
Roads: 12
Stlmts: 3
Cities: 4
Tomm
Points: 4
L. Road
Soldiers: 1
Roads: 8
Game
* It's rennocl's turn to roll. Rolled a 11.
* rennocl gets 1 wheat. Tomm gets 1 wheat.
History
rennocl: you know if you have more than 4, I think you can trade them into the bank,
Dave: yeah but it's not ideal
Tomm: That's true... 4 -> 1
Tomm: Well, I might do, but depends on my roll, I'm afraid
Tomm: Dave: Wheat for a clay?
Dave: sure, can you do 2 for 2, or do you just want 1 for 1
Tomm: just 1-4-1 I'm afraid
Dave: fair enough
Tomm: Oh... now I get wheat(!)
Dave: heh
Chat
Dave
Points: 2
Soldiers: 1
Resources: 5
Dev. Cards: 1
Roads: 11
Stlmts: 3
Cities: 4
Sit Here
Road:
Cost:
Settlement:
Game Options...

Multi party dialogue structures

234	gotwood4sheep	anyone got wheat for a sheep ?
235	inca	sorry, not me
236	CheshireCatGrin	nope. you seem to have lots of sheep !
237	gotwood4sheep	yup baaa
238	dmm	i think i'd rather hang on to my wheat i'm afraid
239	gotwood4sheep	kk I'll take my chances then...



STAC corpus annotations

- much larger corpus than Annodis. $\sim$ 1100 negotiation dialogues.
- refined annotation tool Glozz
- better κ on brute discourse structure measurements (.72 attachment on structures, .58 on relation tags)

Annotation scheme

- segmentation of dialog turns into discourse units
- labelling with domain-related speech acts (negotiation moves)
- relational rhetorical annotation familiar from ANNODIS but with relations for dialogue.

Domain level acts

- offer : *I'll give you 2 clay for a rock*
- counteroffer : *How about 2 clay for a wheat ?*
- accept : *OK, it's a deal.*
- refusal : *I don't think so.*
- has-resource : *I have wheat*
- strategic comment : *joel fancies a bit of your clay*
- other (non relevant for negotiation)

Example annotation

Speaker	Id	Turn	Dom. function	Rhet. fur
Euan	47	[And I alt tab back from the tutorial.]_1	other	
		[What's up ?]_2	other	Result*(4
Joel	48	[do you want to trade ?]	offer ⟨Joel, ?, ?,Euan⟩	Q-elab(4
Card.	49	[joel fancies a bit of your clay]	strat.-comment	Expl*(48
Joel	50	[yes]	other	Ackn(49,
Joel	51	[!]	other	Commen
Euan	52	[Whatcha got ?]	counteroffer ⟨Euan, ?, ?,Joel⟩	Q-elab([4
Joel	53	[wheat]	has-resources ⟨Joel,wheat⟩	QAP(52,
Euan	54	[I can wheat for clay.]	counteroffer ⟨Euan,wheat,clay,Joel⟩	Elab([52,
Joel	55	[awesome]	accept(54)	Ackn(54,

Parsing in multiparty dialogue

- non-treelike structures certainly exist.
- long distance crossing dependencies
- different decoders needed.
- Our strategy : augmenting tree structures, and thinning Maximal Spanning DAGs with constraints. Experimenting with ILP.

Parsing experiments

Method	Attachment			Relations		
	F1	prec	rec	F1	prec	rec
Last	56.2	57.8	54.7	33.6	34.6	32.7
Local	52.9	71.6	42.0	37.2	50.4	29.5
MST	60.0	62.1	58.0	42.0	43.5	40.6
MST+	**60.4**	61.9	59.0	**42.4**	43.4	41.4
MSDAG	20.5	11.5	94.7	12.7	7.1	58.8
ILP	46.8	35.8	67.5	30.7	23.5	44.3

For MST+ we have used a window of 3 and a probability threshold of 0.3

Conclusions

- discourse parsing difficult but extends to dialogue pretty robustly.
- to do better we need better features (still very shallow)
- a better understanding of structures and constraints on DAGs.