Structured Prediction Using Decoding in the Context of Discourse Parsing for Chat Dialogues: From Classical to Neural Approaches

Stergos Afantenos

ENSEEIHT - INP
12 novembre 2018

Introduction

Supervised classification

- We have a set of labeled examples

$$\{\mathbf{x}_i, y_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} P(\mathbf{x}, y) \quad \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}$$

- We need to learn a function $f : \mathcal{X} \mapsto \mathcal{Y}$ that predicts $y = f(\mathbf{x})$ on future data $\mathbf{x}$ with $(\mathbf{x}, y) \overset{i.i.d.}{\sim} P(\mathbf{x}, y)$
- The learned function can be linear $y = \text{argmax}_{y \in Y} \mathbf{w}_y \mathbf{x}$ or non-linear, learned from a neural network.
- Crucially, $\mathcal{Y}$ is a set of usually few classes that are distinct between them.

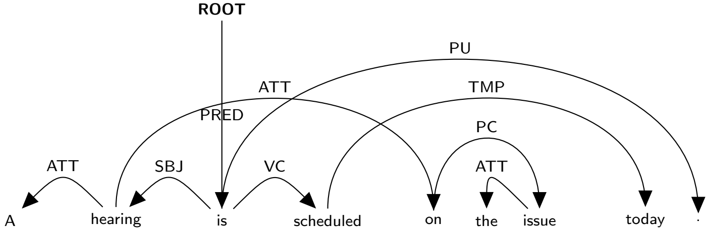
Structured prediction

What happens when $y \in \mathcal{Y}$ is a complex object?

- y can be a sequence

$\mathbf{x} =$	John	saw	Mary	with	the	telescope
$y =$	noun	verb	noun	preposition	article	noun

- y can be a tree or a graph



Structured prediction (cont.)

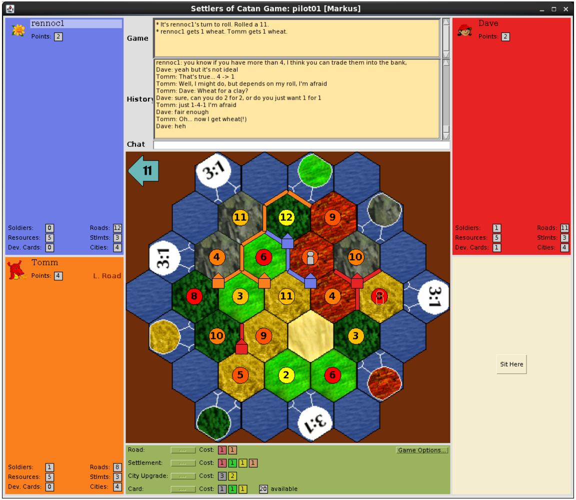
- If we consider y as a separate class for every sequence/tree/graph then we can have exponentially many classes!
- There are various ways to perform structured output prediction:
  - Decoding over a local probability distribution
  - Using the kernel trick in SVMs
  - Using an approach similar to SparseMap (Niculae et al. 2018)

Discourse for multi-party dialogues

- Discourse parsing for monologues has been extensively investigated
- Discourse parsing for other forms of human communication on the other hand has not received the same attention from the computational linguistics community

The Settlers of Catan

A board game where **2-4 players compete** for establishing settlements and roads on an island, gathering and **negotiating** resources in the process.



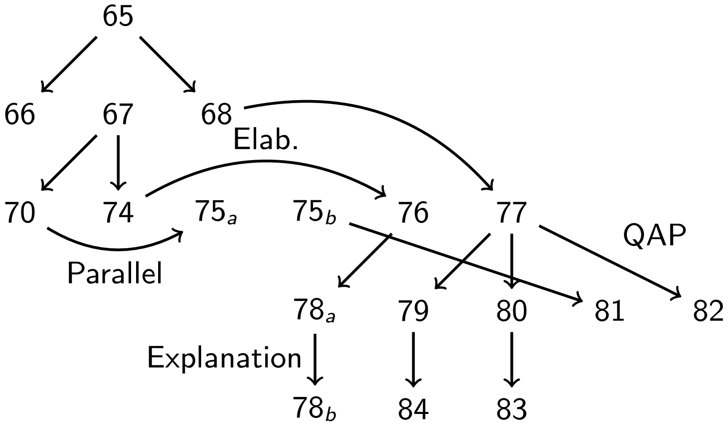
A sample dialogue

Our input

65	lj	anyone want sheep for clay?
66	gw	got none, sorry :(
67	gw	so how do people know about the league?
68	wm	no
70	lj	i did the trials
74	tk	i know about it from my gf
75	gw	[yeah me too,]$_a$
		[are you an Informatics student then, lj?]$_b$
76	tk	did not do the trials
77	wm	has anyone got wood for me?
78	gw	[I did them]$_a$ [because a friend did]$_b$
79	gw	lol william, you cad
80	gw	afraid not :(
81	lj	no, I'm about to start math
82	tk	sry no
83	gw	my single wood is precious
84	wm	what's a cad?

Dependency graph

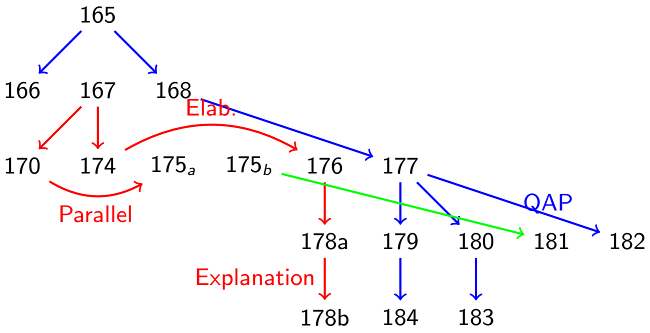
Our target



Concurrent discussions

165	lj	anyone want sheep for clay?
166	gw	got none, sorry :(
167	gw	so how do people know about the league?
168	wm	no
170	lj	i did the trials
174	tk	i know about it from my gf
175	gw	[yeah me too,]$_a$
		[are you an Informatics student then, lj?]$_b$
176	tk	did not do the trials
177	wm	has anyone got wood for me?
178	gw	[I did them]$_a$ [because a friend did]$_b$
179	gw	lol william, you cad
180	gw	afraid not :(
181	lj	no, I'm about to start math
182	tk	sry no
183	gw	my single wood is precious
184	wm	what's a cad?

Concurrent discussions

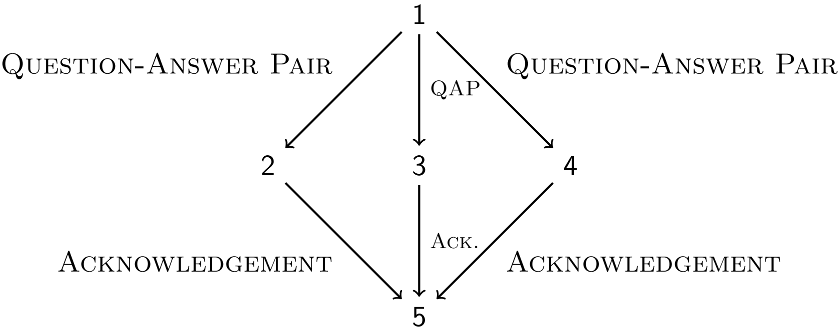


A smaller example

1	Alice	anyone got wheat for a sheep?
2	Bob	sorry, not me
3	Clara	nope. you seem to have lots of sheep!
4	Dan	i think i'd rather hang on to my wheat i'm afraid
5	Alice	kk I'll take my chances then...

A smaller example

1	Alice	anyone got wheat for a sheep?
2	Bob	sorry, not me
3	Clara	nope. you seem to have lots of sheep!
4	Dan	i think i'd rather hang on to my wheat i'm afraid
5	Alice	kk I'll take my chances then...



How can we represent dialogue?

- Penn Discourse TreeBank (PDTB) uses directed edges, without structural constraints

How can we represent dialogue?

- Penn Discourse TreeBank (PDTB) uses directed edges, without structural constraints
- Rhetorical Structure Theory (RST) represents discourse as trees, which is too restrictive

How can we represent dialogue?

- Penn Discourse TreeBank (PDTB) uses directed edges, without structural constraints
- Rhetorical Structure Theory (RST) represents discourse as trees, which is too restrictive
- Segmented Discourse Representation Theory (SDRT) uses two-layered directed acyclic graphs

How can we represent dialogue?

- Penn Discourse TreeBank (PDTB) uses directed edges, without structural constraints
- Rhetorical Structure Theory (RST) represents discourse as trees, which is too restrictive
- Segmented Discourse Representation Theory (SDRT) uses two-layered directed acyclic graphs

SDRT's flexible structure is best suited for chats

SDRT graphs

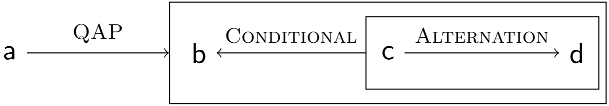
Given a discourse segmented in EDUs, an SDRT graph is a tuple (V, E_1, E_2, ℓ), where

- Vertex set V contains discourse units (DUs)
- Edge set E_1 contains discourse relations between DUs
- Edge set E_2 represents *Complex Discourse Units (CDUs)*
- Function ℓ assigns a label to discourse relation edges

Complex Discourse Units

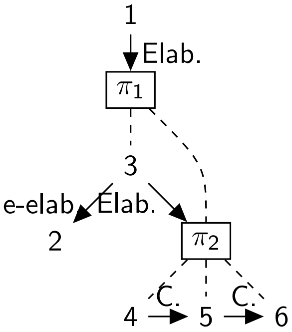
Example

Alice	[Do you have a sheep?]$_a$
Bob	[I do,]$_b$ [if you give me clay]$_c$
Bob	[or wood.]$_d$



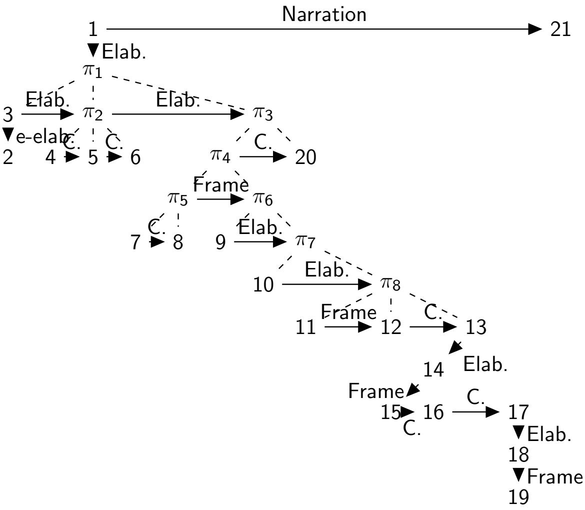
Complex Discourse Units (cont.)

[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



Complex Discourse Units (cont.)

A more complicated example



Complex Discourse Units (cont.)

No reliable method has been identified in the literature for identifying CDUs.

We approximate CDUs in the SDRT hypergraph by relations between EDUs only, thus creating a dependency graph.

Distributing relations

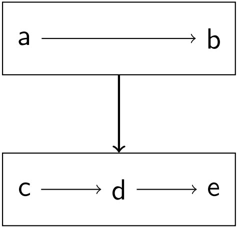
$$\begin{tikzcd} \boxed{a \longrightarrow b} \arrow[d] \\ \boxed{c \longrightarrow d \longrightarrow e} \end{tikzcd}$$

$$\begin{tikzcd} a \arrow[rr] \arrow[d] & & b \\ c \arrow[r] & d \arrow[r] & e \end{tikzcd}$$

No distribution

Head points to head

Distributing relations



[I'll buy a card]$_a$ [and not a road]$_b$ [because I have sheep]$_c$ [and wheat]$_d$ [and ore]$_e$

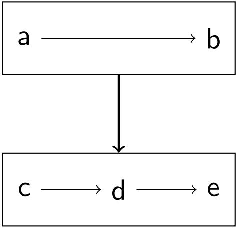
$$\begin{tikzcd}
a \arrow[rr] \arrow[d] & & b \\
c \arrow[r] & d \arrow[r] & e
\end{tikzcd}$$

No distribution
Head points to head

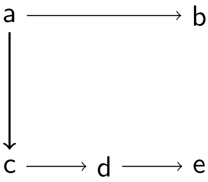
$$\begin{tikzcd}
a \arrow[rr] \arrow[d] \arrow[dr] \arrow[drr] & & b \\
c \arrow[r] & d \arrow[r] & e
\end{tikzcd}$$

Partial distribution
Relation semantics determine distribution to the source/target CDU components

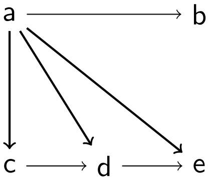
Distributing relations



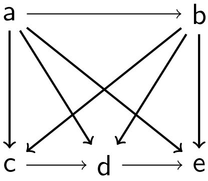
[I'll buy a card]$_a$ [and not a road]$_b$ [because I have sheep]$_c$ [and wheat]$_d$ [and ore]$_e$



No distribution
Head points to head



Partial distribution
Relation semantics determine distribution to the source/target CDU components



Full distribution
All relations distribute to every component

Discourse structure annotation

- 4 naive annotators where involved; they were trained on 22 negotiation dialogues with 560 turns.
- 0.72 kappa on structure and 0.58 kappa on labelling
- Expert annotators adjudicated the naive annotators.
- Adjudication involved five separate phases

Discourse structure annotation

- 4 naive annotators where involved; they were trained on 22 negotiation dialogues with 560 turns.
- 0.72 kappa on structure and 0.58 kappa on labelling
- Expert annotators adjudicated the naive annotators.
- Adjudication involved five separate phases

Dataset overview:

	Total	Training	Testing
Dialogues	1081	965	116
Turns	9160	8166	994
EDUs	10678	9546	1132
Relation instances	10513	9421	1092
CDUs	1284	1132	152

A dialogue includes a negotiation phase during a game

Distribution of annotated relations

	Total	Training	Testing
Comment	1851	1684	167
Clarification_question	260	240	20
Elaboration	869	771	98
Acknowledgment	1010	893	117
Continuation	987	873	114
Explanation	437	407	30
Conditional	124	105	19
Question-answer_pair	2541	2236	305
Alternation	146	128	18
Q-Elab	599	525	74
Result	578	551	27
Background	61	58	3
Narration	130	116	14
Correction	212	189	23
Parallel	215	196	19
Contrast	493	449	44
TOTAL	10513	9421	1092

Learning structures vs Local Models

Ideally:

$$h : \mathcal{X}_{E^n} \mapsto \mathcal{Y}_{\mathcal{G}}$$

Realistically:

$$h : \mathcal{X}_{E^2} \mapsto \mathcal{Y}_{R}$$

Problems with this approach

- We have no guarantees that structures will be well formed
- graphs might be disconnected
- we might have cycles
- the Right Frontier Constraint might not be respected
- etc.

How can we alleviate this problem?

Do structured decoding over local probability distributions

- Maximum Spanning Trees (MST)
- Integer Linear Programming (ILP)

Maximum Spanning Trees (MST)

Local Probability Distributions

We used a regularized Maximum Entropy model:

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

Features used

Category	Description
Positional	Speaker initiated the dialogue
-	First utterance of the speaker in the dialogue
-	Position in dialogue
-	*Distance between EDUs*
-	*EDUs have the same speaker*
Lexical	Ends with exclamation mark
-	Ends with interrogation mark
-	Contains possessive pronouns
-	Contains modal modifiers
-	Contains words in lexicons
-	Contains question words
-	Contains a player's name
-	Contains emoticons
-	First and last words
Parsing	Subject lemmas given by syntactic dependency parsing
-	Dialogue act according to (Cadilhac et al, 2013)

The turn Constraint

- Within a turn people can have a full discourse model, including backward links
- Outside turns, we cannot have backward links

The turn Constraint

- Within a turn people can have a full discourse model, including backward links
- Outside turns, we cannot have backward links

Example:

Although he was very tired, he still came to the meeting.

The turn Constraint

- Within a turn people can have a full discourse model, including backward links
- Outside turns, we cannot have backward links

Example:

Although he was very tired, he still came to the meeting.

- We thus build two different local models applying this constraint
  - Intra-turn: training contains all pairs of EDUs (i, j) with $i \neq j$
  - Inter-turn: training contains all pairs of EDUs (i, j) with $i < j$
- We apply it during decoding also

Decoders

- Baseline decoder (Local)

$$\hat{r} = \operatorname*{argmax}_{r} \left(\frac{1}{Z(c)} \exp \left(\sum_{i=1}^{m} w_i f_i(p, r) \right) \right)$$

- Maximum Spanning Trees (MST)

$$T^* = \operatorname*{argmax}_{T \text{ a spanning tree of } G} \sum_{e \in E(T)} w(e)$$

$$w(e) = \log \left(\frac{p(e)}{1 - p(e)} \right)$$

Evaluation F1 scores on test corpus

Method	Unlabelled	Labelled
Last	0.584	0.391
Local	0.483	0.429
MST	0.671	0.516

Integer Linear Programming (ILP)

Integer Linear Programming: an introduction

We define an optimization problem where all variables are integers:

$$\begin{aligned} \text{maximize} \quad & c^T x \\ \text{subject to} \quad & Ax \leq b \\ & x \geq 0 \\ \text{and} \quad & x \in \mathbb{Z}^n \end{aligned}$$

- Structural freedom
- Easy to parametrize
- Versatile constraints on need

Our model

Pair modelization: Maximum Entropy model

The model provides us with two real-valued functions:

$$\begin{aligned} s_a : &\quad [1..n]^2 \mapsto [0,1] \\ s_r : &\quad [1..n]^2 \times [1..m] \mapsto [0,1] \end{aligned}$$

Graph building: Integer Linear Programming

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_i \sum_j \left(a_{ij} s_a(i,j) + \sum_k r_{ijk} s_r(i,j,k) \right) \\ \text{subject to} & \text{our set of constraints} \end{array}$$

Structural constraints

- Acyclicity
- Unique root
- Connectedness
- Turn Constraint

Edge count bounds

Outgoing edge cap

An utterance can elicit a limited number of reactions:

$$\forall i \quad \sum_j a_{ij} \leq \omega$$

Edge count bounds

Outgoing edge cap

An utterance can elicit a limited number of reactions:

$$\forall i \quad \sum_j a_{ij} \leq \omega$$

Density cap

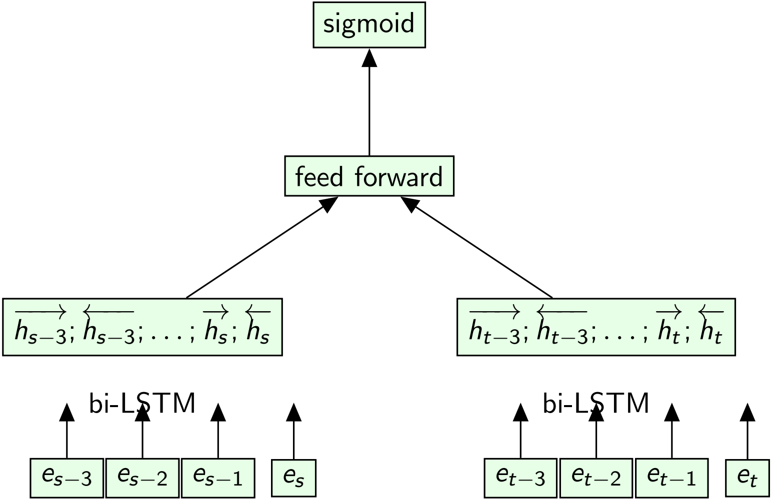
An unbounded number of edges would result in a near-complete graph, as the objective function is increasing.

$$\sum_i \sum_j a_{ij} \leq \delta(n-1)$$

Evaluation F1 scores on test corpus

Method	Unlabelled	Labelled	Edge count
No distribution			10191
Last	0.584	0.391	
Local	0.483	0.429	
MST	0.671	0.516	
ILP	**0.689**	**0.531**	
Partial distribution			11734
Last	0.593	0.426	
Local	0.471	0.396	
MST	0.647	0.488	
ILP	**0.668**	**0.519**	
Full distribution			13675
Last	0.582	0.420	
Local	0.541	0.443	
MST	0.613	0.466	
ILP	**0.675**	**0.527**	

Neural network architecture



Evaluation

	Bi-LSTM			MST Decoding		
	P	R	F1	P	R	F1
LAST	50.26%	64.31%	56.42%	-	-	-
Distance 1	9.10%	18.24%	12.14%	14.88%	19.04%	16.70%
Distance 2	52.49%	57.58%	54.92%	50.98%	65.22%	57.22%
Distance 3	52.12%	62.82%	56.98%	51.60%	66.02%	57.92%
Distance 4	57.35%	55.98%	56.66%	52.22%	66.81%	**58.62%**

Future work

- Use the learned representations as input to an SVN structured prediction framework (joint work with Phuong Nguyen, Edouard Pauwels and Mathieu Serrurier)
- Disentangle threads of conversations.
- Perform semi-supervised learning (joint work with Luce Le Gorrec and Sandrine Mouysset)