L'analyse discursive automatique face au manque de données

Séminaire CLLE - 07/12/2020

Chloé Braud - CNRS - IRIT (Toulouse) - chloe.braud@irit.fr

Travaux en collaboration avec Philippe Muller, Mathieu Morey, Charlotte Roze, Pascal Denis, Mikel Iruskieta, Maximin Coavoux, Anders Sogaard et Ophélie Lacroix.

Whenever we read something closely, with even a bit of sensitivity, ***text structure leaps off the page at us****. We begin to see elaborations, explanations, parallelisms, contrasts, temporal sequencing, and so on. These relations bind contiguous segments of text into a* ***global structure for the text as a whole****.*

Hobbs, 1985

1. L’analyse discursive, en théorie et en pratique
2. Segmentation discursive
3. Identification des relations discursives
4. Parsing discursif
5. Bilan, évaluation et conclusions

→ Approches : gérer le manque de données

→ Questions :

- **type d'information linguistique nécessaire pour une tâche discursive**
- **nature des relations discursives**
- **similarités entre les langues pour ce niveau d'analyse**
- **(utilité de l’analyse discursive en TAL)**

Analyse discursive

Théorie et pratique

- **Analyse discursive**
  - **Cohérence et cohésion**
  - **Relations discursives**
- **Cadres théoriques / schémas d'annotation**
  - **Consensus partiel**
  - **Les corpus**
- **Systèmes automatiques**
  - **(RST) parsing**
  - **(PDTB) chunking**

Analyse discursive

- Phénomènes qui couvrent plusieurs phrases :
  - Topics (topic segmentation),
  - Liens temporels
  - Anaphores et co-référence
  - Relations discursives / rhétoriques
- Structure discursive :
  - révéler la cohérence textuelle
  - interpréter des documents (i.e. inférences sur le contenu)
- Liens entre les formes d'organisation textuelle, e.g. :
  - contraintes discours / temporel : e.g. often the effect after the cause
  - discours / topic : e.g. some relations require to keep the same topic
  - discours / coreference : e.g. some relations block a potential referent

Cohérence et cohésion

- Document: pas une séquence aléatoire de phrases
  - A text is **cohesive** if its elements are linked together (non structural textual relations)
  - A text is **coherent** if it makes sense (structural relation between segments).

→ Document = un groupe de phrases cohérent et structuré

Exemples:

1) Paul fell. Marie pushed him. (Cause)
2) Paul fell. He likes spinach. (??)
3) Paul went to Istanbul. He has to attend a conference. (Reason)
4) Paul went to Istanbul. He likes spinach. (? Reason)
5) *It's like going to disney world for car buyers. I have to say that Carmax rocks. We bought **it** at Carmax, and I continue to have nothing bad to say about that company. After our last big car milestone, we've had an odyssey with cars. [Résumé automatique, Mithun and Kosseim, 2011]

Relations discursives

Les connecteurs discursifs :

- contribuent à la cohésion et à la cohérence
- explicitent la relation entre les unités de texte adjacentes
  - ***cependant*** : signale une relation contrastive
  - ***de plus*** : indique que le texte qui suit élabore ou renforce le point présenté juste avant,
  - ***pendant ce temps*** : indique que deux événements sont contemporains
  - ***si...alors*** : indique une relation conditionnelle.

Lexiques de connecteurs disponibles pour différentes langues

Relations discursives

Les relations discursives lient :

- le contenu sémantique de deux unités (1)
- ou un acte de parole et le contenu sémantique d'une autre unité (2)
- on retrouve ces relations sans la présence de connecteur = implicite (2)
- on retrouve ces phénomènes entre phrases et à l'intérieur des phrases (3)

1) This cute child turns out to be a blessing and a curse. She gives the Artist a sense of purpose, **but** also alerts him to the serious inadequacy of his vagrant life. **(Cause-reason)**
2) Mrs. Yeargin is lying. They found students (..) who said she gave them similar help. **(Pragmatic Cause-justification)**
3) Typically, money-fund yields beat comparable short-term investments **because** portfolio managers can vary maturities and go after highest rates. **(Cause-reason)**

(PDTB)

Cadres théoriques - consensus partiel

- Niveau d'analyse : le document
- Elementary Discourse Unit (EDU): clauses (en général), au plus une phrase
- Relations discursives :
  - sémantico-pragmatiques
  - binaires (en général)
  - inter- ou intra-phrastiques
  - Explicite ou implicite : présence/absence d'un connecteur discursif

1) La chouette hulotte est un animal nocturne, **mais** elle peut vivre le jour.
2) Les tours se sont effondrées moins de deux heures plus tard (Result) **entraînant** l'immeuble du Marriott World Trade Center dans leur chute. (Sequence) La tour 7 du WTC s'est effondrée dans l'après-midi **en raison d'**incendies et des dégâts occasionnés par la chute des Twin Towers. (Annodis)

Cadres théoriques / schémas d'annotation : 2 vues

Structure discursive hiérarchique (RST, SDRT, DLTAG, GraphBank…)

- Structure : arbres/graphes couvrant les documents
- Tente de donner une interprétation aux documents
- Annotation difficile !

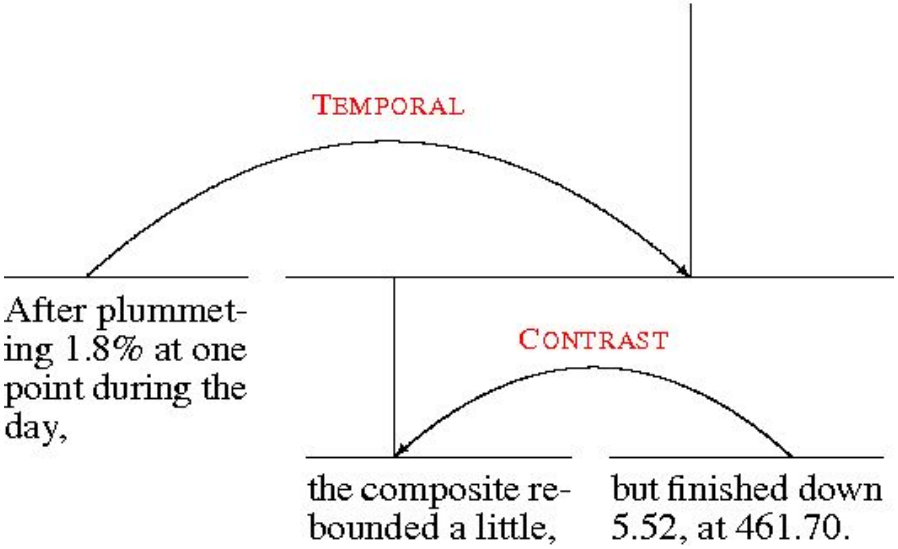
Cohérence locale (PDTB)

- "theory neutral": fondé sur le lexique
- Structure plate, pas de couverture totale des documents
- Accords inter-annotateurs plus élevés (corpus plus larges)

Rhetorical Structure Theory [Mann and Thompson, 1988]

RST DT [carlson et al., 2001] https://www.sfu.ca/rst/

- Relations**: intentions de l'auteur**
- Distinction relation/segments nucleus-satellite / multi-nucléaire
- 1 relation par paire de segments
- Structure: **arbres couvrant** les documents
- **78 relations, 16 classes**
- **385 documents** - Wall Street Journal articles
- Le plus utilisé parce que : **Arbres** !!
  - relations étranges : *attribution, same-unit*...
  - segmentation étrange
  - définition des relations difficile à apréhender (https://www.sfu.ca/rst/01intro/definitions.html)



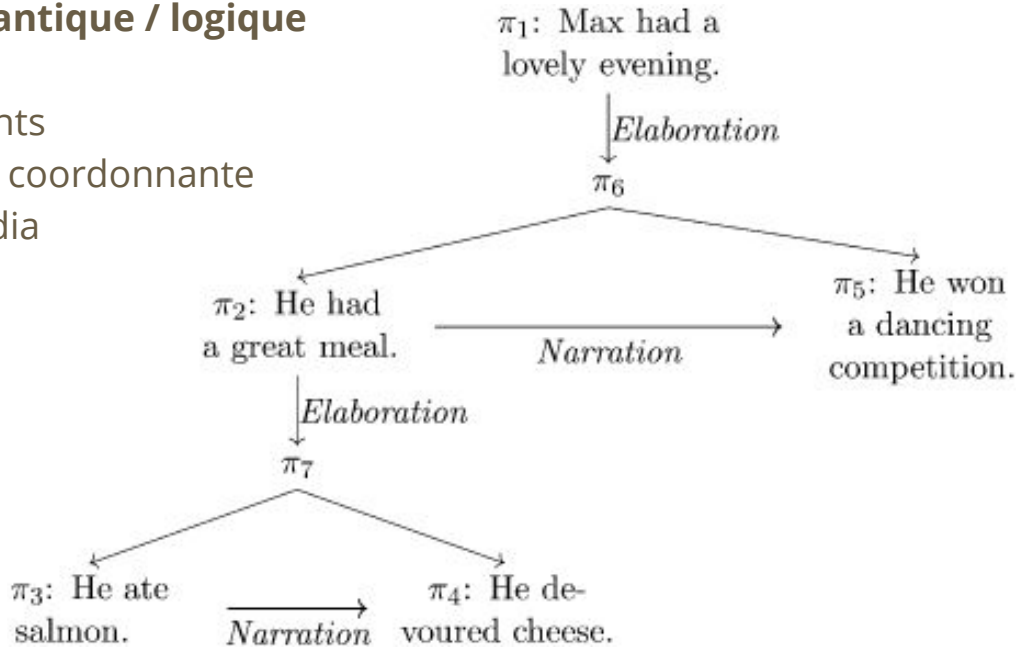
RST relation set (relations and classes)

- **Attribution**: attribution, attribution-negative
- **Background**: background, circumstance
- **Cause**: cause, result, consequence
- **Comparison**: comparison, preference, analogy, proportion
- **Condition**: condition, hypothetical, contingency, otherwise
- **Contrast**: contrast, concession, antithesis
- **Elaboration**: elaboration-additional, elaboration-general-specific, elaboration-part-whole, elaboration-process-step, elaboration-object-attribute, elaboration-set-member, example, definition
- **Enablement**: purpose, enablement
- **Evaluation**: evaluation, interpretation, conclusion, comment
- **Explanation**: evidence, explanation-argumentative, reason
- **Joint**: list, disjunction
- **Manner-Means**: manner, means
- **Topic-Comment**: problem-solution, question-answer, statement-response, topic-comment, comment-topic, rhetorical-question
- **Summary**: summary, restatement
- **Temporal**: temporal-before, temporal-after, temporal-same-time, sequence, inverted sequence
- **Topic Change****: topic-shift, topic-drift

Segmented Discourse Representation Theory [Asher and Lascarides 2003]

Annodis [Afantenos et al. 2012]

- Relations : définitions fondées sur la **sémantique / logique**
- Relations multiples
- Structure: **graphes couvrant** les documents
- **18 relations,** distinction subordonnante / coordonnante
- **86 documents -** Est Républicain + Wikipédia
- unitées enchâssées



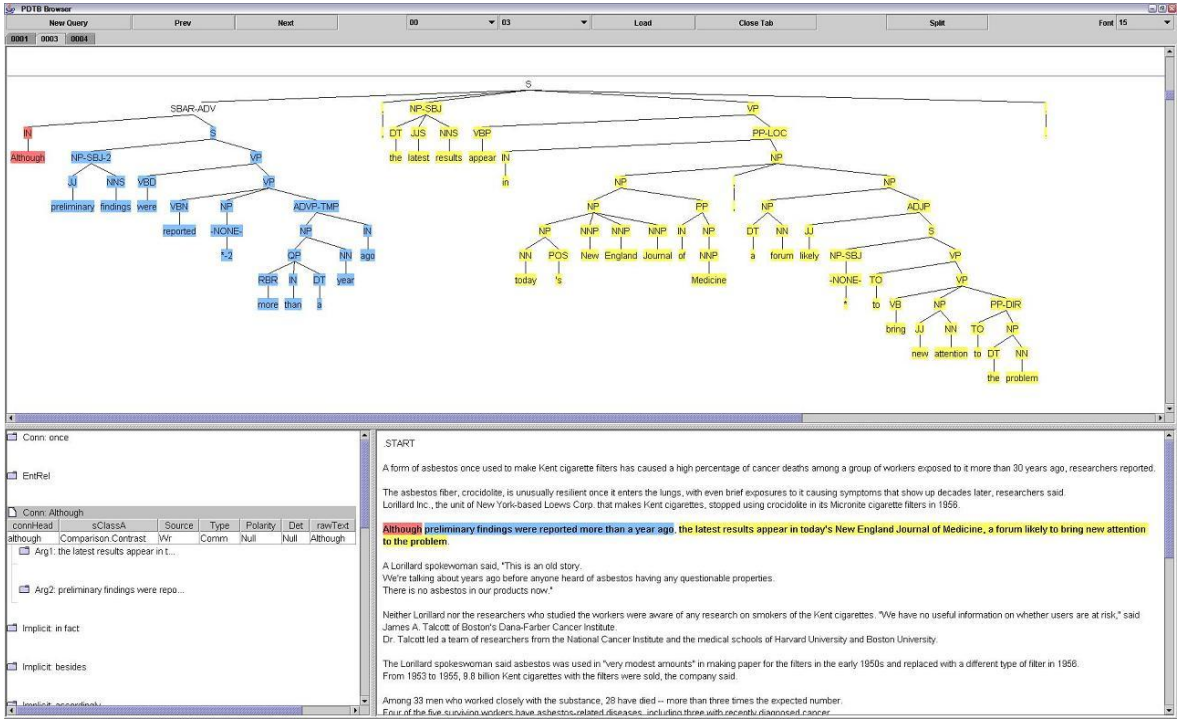
STAC [Asher et al. 2016]

- Conversations stratégiques
- Chats multi-parties (jeu de plateau)
- 45 conversations

Penn Discourse TreeBank

PDTB 2.0 [Prasad et al., 2008]

- Annotation basée sur les **connecteurs et l'adjacence**
- Relations multiples (2% implicites)
- Distinction Arg1 / Arg2 (=conn)
- **Plat/pas de structure**
- Hiérarchie de relations
  - **4 classes,**
  - **16 types,**
  - **23 sous-types**
- **2 259 WSJ articles (<PTB)**



Penn Discourse TreeBank 2.0

- 2 259 documents, 40 600 annotations :
  - Relations **explicites = 18,459**
    - le connecteur + les arguments + jusqu'à 2 sens
    - Connecteurs en liste fermée = 100
    - Arguments = principe de minimalité
  - Relations **implicites = 16,224**
    - phrases adjacentes + up to 2 connectives)
  - **Lexicalisations alternatives = 624**
    - similaire aux implicites mais impossible d'insérer un connecteur (redondance)
  - **Relations d'entité : 5,210**
  - **No Relation: 254**

|Mrs Yeargin is lying.|1 **Implicit = because** |They found students in an advanced class a year earlier who said she gave them similar help.|2 (CONTINGENCY:Pragmatic Cause:justification)

PDTB relation set

- TEMPORAL
  - Asynchronous
  - Synchronous: precedence, succession
- CONTINGENCY
  - Cause: result, reason
  - Pragmatic cause: justification
  - Condition: hypothetical, general, unreal present, unreal past, real present, real past
  - Pragmatic condition: relevance, implicit assertion
- COMPARISON
  - Contrast: juxtaposition, opposition
  - Pragmatic contrast
  - Concession: expectation, contra-expectation
  - Pragmatic concession
- EXPANSION
  - Conjunction
  - Instantiation
  - Restatement: specification, equivalence, generalization
  - Alternative: conjunctive, disjunctive, chosen alternative
  - Exception
  - List

Les corpus

+**SciDTB**: Scientific Abstracts (2018)
+**Turnitin**: student essays (2019)
+**Molweni**: Multiparty Dialogues (2020)

+**TED talks**: English, Polish, German, Russian, European Portuguese, and Turkish (2019)

Language	RST	SDRT	PDTB	Other
English	RST DT	DisCor	PDTB	GraphBank
	Instructional	STAC corpus	Biomedical DRB	-
	SFU Review Corpus	-	-	-
English, Spanish, Basque	Multilingual RST DT	-	-	-
Spanish	Spanish RST DT	-	-	-
Italian	-	-	LUNA	-
Fraench	-	ANNODIS	French DT	-
German	PCC	-	(PCC)	Tüba-D/Z
	Twitter?	-	-	-
Dutch	Dutch RST DT	-	-	DiscAn
Danish	-	-	-	Copenhagen DeT
Arabic	-	Arabic DT	Leeds Arabic DT	-
Basque	Basque RST DT	-	-	-
Catalan	-	-	-	CatDiG
Portuguese	Rhetalho	-	-	-
Brazilian	CSTNews	-	-	-
Czech	-	-	Prague DT	-
Turkish	-	-	Turkish DT	-
Chinese	Chinese RST DT	-	Chinese DT	-
Hindi	-	-	Hindi DRB	-
Tamil	?	-	-	-

DT = Discourse Treebank ; DRB = Discourse Relation Bank

L’analyse discursive automatique

- beaucoup d’applications de TAL se limitent au niveau de la phrase
  - mais problématique, cf exemple du résumé automatique
- l’information discursive est nécessaire (si nécessaire) à la fin de la chaîne
- signifie aussi : besoin des informations des étapes précédentes et propagation d’erreurs

Les tâches

- RST/SDRT = discourse parsing
- PDTB = discourse chunking (shallow discourse parsing)

Parsing discursif (RST ou SDRT)

Etapes :

1. segmentation en Elementary Discourse Units (EDU)
2. processus récursif : construction de la structure
    - attachement
    - identification de la relation
    - + bonus RST : étiquetage des segments entre nucleus ou satellite

Parsing discursif (RST ou SDRT)

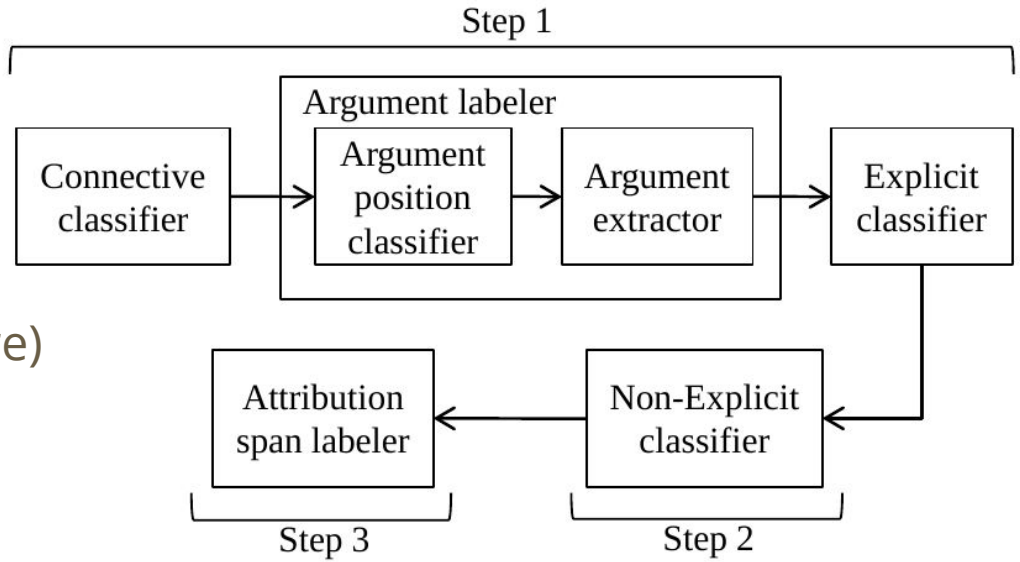
Etapes :

1. **segmentation** en Elementary Discourse Units (EDU)
2. processus récursif : **construction de la structure**
    - attachement
    - identification de la relation
    - + bonus RST : étiquetage des segments entre nucleus ou satellite

Chunking discursif ou shallow parsing (PDTB)

Pipeline :

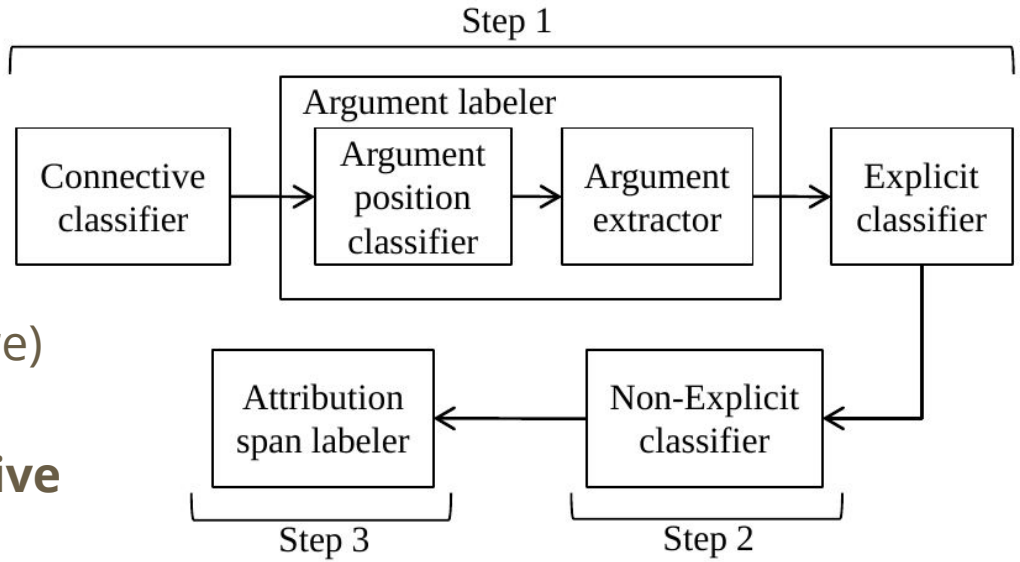
1. Identifier les connecteurs
2. Identifier leurs arguments
   a. position (inter / intra, ordre)
   b. frontières
3. Identifier la relation discursive
   a. explicites
   b. implicites
   c. (entRel, AltLex, NoRel)



Chunking discursif ou shallow parsing (PDTB)

Pipeline :

1. Identifier les connecteurs
2. Identifier leurs arguments
   a. position (inter / intra, ordre)
   b. frontières
3. **Identifier la relation discursive**
   a. explicites
   b. implicites
   c. (entRel, AltLex, NoRel)



Différentes applications

- Co-référence [Cristea et al., 1999]
- Automatic summarization [Sporleder and Lapata, 2005; Gerani et al., 2014; Schrimpf, 2018, Koto et al, 2019]
- Question Answering [Verberne, 2007]
- Sentiment analysis [Bhatia et al., 2015; Lee et al. 2018]
- Essay scoring [Higgins et al., 2004; Mesgar and Strube, 2018]
- Summary coherence rating [Nguyen and Joty, 2017], coherence Modeling [Li and Jurafsky, 2017; Mesgar and Strube, 2018]
- Machine translation [Meyer and Webber, 2013; Born et al., 2017]
- Popularity prediction [Koto et al, 2019]
- Sarcasm detection [Lee et al. 2018]
- Alzheimer analysis [Abdalla et al. 2018]

Travaux présentés

Gérer le manque de données

- **Segmentation**
  - sans oracle
  - en cross-lingue
  - embeddings only
- **Relations discursives**
  - explicites vs implicites
  - décomposition en primitives
- **RST parsing**
  - en cross-lingue
  - en multi-tâche

Segmentation discursive

Segmentation discursive

- Considérée comme une tâche facile et résolue
- Mais :
  - quasi uniquement pour l'anglais, les monologues et le WSJ, pas de comparaisons entre langues, modalités, cadres théoriques…
  - nécessite une segmentation gold en phrase (pas si facile que ça …)
  - nécessite une analyse syntaxique (semble cher pour une simple 'tokenisation')
- D'où :
  - Cross-lingual and cross-domains discourse segmentation. (2017)
  - Does syntax helps discourse segmentation? Not so much. (2017)
  - ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents (2019)

Segmentation discursive - RST DT

But : segmenter un document en EDUs

- clauses, max phrase, mais RST DT = grain fin cf les règles tagging manual :
  - Includes both speech acts and other cognitive acts:
    - (a) |The company **says**| |it will shut down its plant.|
    - (b) |HDTV is where everybody is going,| |**says** ...|
  - But if the complement is a to-infinitival, do not segment:
    - (c) |The company wants **to** shut down its plant.|
  - But segment infinitive clause marking a purpose relation:
    - (d) |A grand jury has been investigating whether officials (...) conspired **to** cover up their accounting| |**to** evade federal income taxes.|
- Autres cas :
  - (e) |Mr. Volk, 55 years old, succeeds Duncan Dwight,| |**who** retired in September.|
  - (f) |The Tass news agency said the 1990 budget anticipates income of 429.9 billion rubles| |**($US693.4 billion)**|

Segmentation discursive sans oracle

Idée : problème résolu ? state-of-the-art: 93.7% [Xuan Bach et al, 2012]

- La plupart des travaux sur l'anglais et le WSJ
  - → Résultats sur **4 langues et 3 domaines**
- Segmentation en phrases gold
  - → Segmentation intra-phrastique *vs* **niveau du document**
- Segmentation fondée sur la syntaxe : utilisation des arbres PTB
  - → Peut-on utiliser des **infos UD**? Peut-on s**e passer de syntaxe** ?
- Tokenisation gold
  - → Pré-traitements entièrement **prédits**

[Braud et al. 2017 a,b]

Système

- Réseau neuronal pour prédiction de séquence (binaire niveau du mot)
  - (a) |Texaco acquired Tana] [before it completed those sales .]
  - (b) Texaco_B acquired_I Tana_I before_B it_I completed_I those_I sales_I ._I
- Architecture neuronale (Dynet):
  - stacked k-layer bi-LSTM
  - left and right context: e.g. segment coordinated VPs but not coordinated NPs

Note : Segments enchâssés traités en 3 segments (prob. non optimal)

- (e) [But maintaining the key components (. . .)]_1 **[a stable exchange rate and high levels of imports –]_2** [will consume enormous amounts (. . .).]_3

Traits

- Mots : embeddings initialisés aléatoirement et entraînables
- POS tags (< PTB gold, PTB pred ou UD)
- Syntaxe : supertags fondés sur les arbres UD [Ouchi et al. 2016]
  - Label of incoming dependency
  - Label of the incoming dependency of the head
  - Direction of the incoming dependency
  - POS (UD) of the head
  - Token of the head
  - Token of the head of the head
  - POS + incoming label of the left/right siblings

Erreurs sur mots outil (''to”, ''and”) = feuilles dans les arbres en dépendances

(a) [they parcel out money] [so that their clients can find temporary living quarters,] [buy food] (...) **[and** replaster walls.]

(b) [Under Superfund, those] [who owned, generated **or** transported hazardous waste] [are liable for its cleanup, (. . .)]

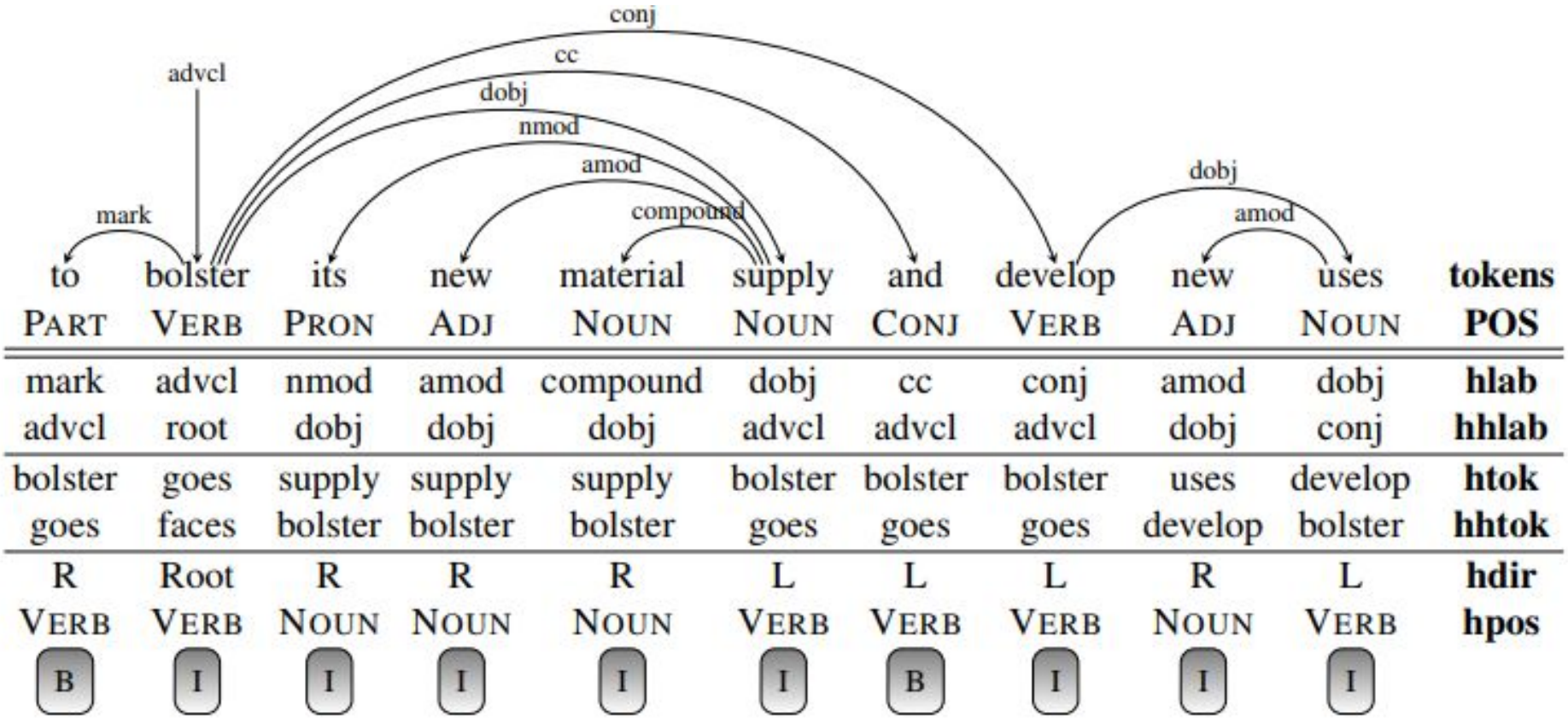


Figure 1: Features extracted from a (part of a) sentence and its predicted UD dependency tree.

Traits simples (intra, En)

- **SOA** : POS+syntax pred = baisse (Stanford parser < PTB)
- **Words** : pas si éparpillés
  - 8,5% peuvent être frontière
  - 1 409 / 16 577 tokens sont des frontières, 104 le sont plus de 10 fois représentant 79,7% des frontières
- **POS >= Words si grain fin**
  - 99.7% PTB POS peuvent être frontière
- **POS UD** manque certaines distinctions
- STAGS ≃ Words / POS PTB pred
- STAGS > POS UD

System	(Morpho-)syntax	F1
[Xuan Bach et al, 2012]	Gold	93.7
[Xuan Bach et al, 2012]	Pred	91.0
Words	-	**81.3**
POS PTB	Gold	**85.0**
POS PTB	Pred	81.6
POS UD	Pred	**76.2**
STAGS	Pred	81.0

Traits combinés (intra, En)

- Words+POS : complémentaire (**esp. UD**)
- **Information syntaxique (bruitée)** : n'aide pas beaucoup
  - POS UD+STAGS > POS UD
    - mais < STAGS
  - Words+POS UD+STAGS < **Words+POS PTB**

Tokenisation prédite :

- Words : +1.4%
- POS UD : -2.2%

System	(Morpho-)syntax	F1
[Xuan Bach et al, 2012]	Gold	**93.7**
[Xuan Bach et al, 2012]	Pred	91.0
Words	-	**81.3**
POS PTB	Gold	85.0
POS PTB	Pred	81.6
POS UD	Pred	76.2
STAGS	Pred	81.0
Words+POS PTB	Gold	**91.0**
Words+POS PTB	Pred	87.6
Words+POS UD	Pred	**87.4**
POS UD+STAGS	Pred	79.6
Words+UD+STAGS	Pred	**86.1**

Résultats niveau du document

Corpus	#Doc	#EDU	#Sent	#Word
En-SFU-DT	400	28,260	16,827	328,362
En-DT	385	21,789	9,074	210,584
Pt-DT	330	12,594	4,385	136,346
Es-DT	266	3,325	1,816	57,768
En-Instr-DT	176	5,754	3,090	56,197
De-DT	174	2,979	1,805	33,591

	Words+UD	Words+UD+STAGS
En (news)	**89.0**	87.0
En-SFU	**87.6**	86.0
En-Instr	**88.3**	86.4
Pt	82.9	**83.0**
Es	**78.7**	78.3
De	85.8	**86.2**

Finalement :

- Syntax doesn't help so much
- Systèmes entièrement prédits disponibles pour plusieurs langues

ToNy: Toulouse Nancy discourse segmenter (2019)

DisRPT shared task:

- 15 corpus
- 10 langues
- styles divers (dont conversations)
- 3 formalismes: RST, PDTB, SDRT
- tout sur github (ou presque)

corpus	lang	framework	train_toks	train_sents	train_docs
deu.rst.pcc	deu	rst	26,831	1,773	142
eng.pdtb.pdtb	eng	pdtb	1,061,222	44,563	1,992
eng.rst.gum	eng	rst	67,098	3,600	78
eng.rst.rstdt	eng	rst	166,849	6,672	309
eng.sdrt.stac	eng	sdrt	36,445	7,689	29
eus.rst.ert	eus	rst	21,122	990	84
fra.sdrt.annodis	fra	sdrt	22,278	880	64
nld.rst.nldt	nld	rst	17,566	1,202	56
por.rst.cstn	por	rst	44,808	1,595	110
rus.rst.rrt	rus	rst	214,484	9,859	140
spa.rst.rststb	spa	rst	43,034	1,577	203
spa.rst.sctb	spa	rst	10,249	304	32
tur.pdtb.tdb	tur	pdtb	398,203	25,080	159
zho.pdtb.cdtb	zho	pdtb	52,061	2,049	125
zho.rst.sctb	zho	rst	8,960	344	32

Un format unique

[**President** Bush insists] [**it** would be a great tool] [**for** curbing the budget deficit] [**and** slicing the lard out of government programs.]

- format begin-inside-outside
- permet une analyse en séquence directement
- **deux versions**:
    - juste tokenisé
    - ou (découpage en phrase donné + analyse syntaxique automatique)
- a dû tordre un peu le bras de certains formalismes (SDRT)

President	**BeginSeg=Yes**
Bush	_
insists	_
it	**BeginSeg=Yes**
would	_
be	_
a	_
great	_
tool	_
for	**BeginSeg=Yes**
curbing	_
the	_
budget	_
deficit	_
and	**BeginSeg=Yes**
slicing	_
(...)	

Quelques questions …

- a-t-on vraiment besoin du découpage en phrases ?
- quelle information est nécessaire/suffisante pour la segmentation ?
  - en particulier a-t-on besoin de plus que l'information lexicale / contextuelle (BERT)
- comparaisons intra formalismes / intra-langues ?
- (comparaisons inter formalismes / intra-langues ?)

Modèle supervisé “simple”

- information en entrée = **embedding du mot** (caractères + mot)
  - fournit robustesse sur la graphie
  - selon modèle par mot : **influence contextuelle** de la phrase +/-
    Random / Glove / BeRT Multi-lingue / BeRT Mono-lingue / ELMo
- modèle séquentiel archi-classique: LSTM bidirectionnel
  - apprend influence spécifique du contexte sur la tâche
  - vers l’avant ou vers l’arrière
- sortie directe sur les labels, pas de modèle de dépendance entre labels
- En pratique: recyclage d’un modèle de NER de AllenNLP

Comparaison sur les corpus anglais

	Rand 50d	GloVe 50d	BERT-E	BERT-M	ELMo
PDTB	77.08	65.17	**90.83**	**89.89**	88.40
GUM	80.58	78.28	86.29	**87.27**	**87.65**
RST DT	78.97	83.21	94.41	**93.72**	**94.75**
STAC	77.43	71.70	84.65	**84.45**	**86.06**

- BERT multilingue quasi au niveau du monolingue sur l'anglais
  - pourquoi s'embêter + délais courts pour la campagne -> on le garde
- BERT limité sur la taille de la phrase : 512 (sous)tokens
  - pas grave sur données avec segmentation en phrase
  - nécessite prédécoupage sinon
  - problème avec certains corpus: phrases trop longues (russe, turc), prétraitements spécifiques

Résultats finaux : https://sites.google.com/view/disrpt2019/shared-task?authuser=0

Transfert cross-domaine

Comparaison de transfert entre corpus même langue et même formalisme

- RST DT = articles de journaux
- GUM = genres mélangés (news, académique, opinion, voyage, interviews, bio, fiction)

Train / Test	RST DT	GUM
RST DT	93	73
GUM	66	96

Finalement :

- pouvoir des embeddings contextuels
- questionnements sur les fondements des corpus discursifs : pourquoi un transfert est si difficile ?
- analyse des erreurs encore à faire

Identification des relations discursives

Identification des relations (implicites)

- Implicites : considérée comme une tâche très difficile car :
  - complexe : informations à plein de niveaux (lexique, syntaxe, sémantique, connaissances du monde) + interactions entre les paires de segments / avec le reste du document
  - mais ‘peu’ de données (environ 18 000 annotations dans le PDTB)
  - surtout si on considère toutes les relations (distributions très déséquilibrée) → mais en fait scores très bas avec 4 classes ...
- Plein de propositions, notamment avec apprentissage par transfert :
  - Combining Natural and Artificial Examples to Improve Implicit Discourse Relation Identification (2014)
  - Comparing Word Representations for Implicit Discourse Relation Classification (2015)
  - Which aspects of discourse relations are hard to learn? Primitive decomposition for discourse relation classification (2019)

Combining natural and artificial examples

Stratégie [Marcu and Echihabi, 2002]

- Explicites et implicites : même ensemble de relations, objets similaires
- Connecteurs peu ambigus : annotation automatique d'explicites
- Supprimer le connecteur = exemple implicite artificiel

Hypothèses :

- Redondance du connecteur
- Similarité des données

(a) Paul fell **(because)** Mary pushed him.

(b) La chouette hulotte est un rapace nocturne, **(mais / *ø)** elle peut vivre le jour.

Problèmes de la méthode

- **Supprimer le connecteur** :
  - Incohérent / agrammatical
  - Modification de la relation (a,b)
- **Annotation automatique :**
  - Erreur sur les arguments
  - Erreur sur l'étiquetage de la relation

→Données dissimilaires : adaptation "de domaine"

(a) **(Although - Contrast / ⌀ - Continuation)** the electronics industry has changed greatly, possibly the greatest change is that very little component level manufacture is done in this country. [Sporleder and Lascarides, 2008]

(b) [Les Amorrites deviennent à la période suivante de sérieux adversaires des souverains d'Ur,] [**(puisqu' - Explanation / ⌀ - Result)** ils commencent alors à migrer en grand nombre vers la Mésopotamie.] (Annodis)

Stratégies (simples)

- Combiner les données
  - union des corpus, avec ou sans pondération
- Combiner les modèles
  - traits représentant les prédictions, interpolation linéaire
- Sélection automatique des exemples artificiels
  - garder les exemples prédits avec une probabilité haute

Données

- Données naturelles :
  - PDTB et Annodis
- Données artificielles
  - automatique : Bllip et Est Républicain
  - manuelle : PDTB explicite

En	Implicit	Artificial - PDTB	Artificial - BLLIP
Temporal	826	3 440	783 080
Contingency	4 185	3 250	315 343
Comparison	2 441	5 471	1 148 245
Expansion	8 601	6 928	715 878
Total	16 053	18 459	2 962 546

Systèmes de référence

	Nat-Only	Art-Only PDTB		Art-Only BLLIP	
Test	Nat	Nat	Art	Nat	Art
Macro F1	39.9	29.2	48.2	26.5	60.1

- relativement haut quand train sur Art et test sur Art = **redondance**
- baisse importante quand train sur Art et test sur Nat = **dissimilarité**

Systèmes avec adaptation

	PDTB		BLLIP	
	no selec	selec	no selec	**selec**
Art-Only	29.2	32.0	26.5	**30.3**
Union	39.3	40.4	31.7	**37.9**
Best	41.2	41.5	39.9	39.9

- importance de la **sélection** surtout pour BLLIP
- problème de bruit : BLLIP < PDTB bien que beaucoup plus de données

Finalement :

- Meilleur système : avec traits lexicaux + autres traits classiques
- Rutherford and Xue : définition de filtres fondés sur la distribution des connecteurs implicites
- D'autres approches depuis : multi-tâche, adaptation de domaine … mais pas clairement résolu

	Best Adapt		**[Rutherford and Xue, 2015]**	
	P	F1	P	F1
Temporal	22.0	**28.0**	38.5	14.7
Contingency	44.9	**45.6**	49.3	43.9
Comparison	44.4	**39.5**	44.9	34.2
Expansion	63.9	62.1	61.4	**69.1**
Macro F1	43.8		38.4	

Décomposition des relations de discours [Roze et al. 2019]

- plusieurs cadres théoriques avec des représentations différentes
- plusieurs schémas d'annotation / corpus
- notamment : pas de consensus sur l'ensemble de relations
  - différents niveaux de granularité, e.g. : contrast (SDRT) = Antithesis, Concession, Contrast (RST)
- pourtant encodage des mêmes informations sémantiques et pragmatiques

→ **avoir une représentation qui rend explicite cette info commune ?**

- résultats bas pour les implicites malgré une grande variété d'approches

→ est-ce que le problème vient plutôt de la représentation des données ou de la **façon dont on modélise la tâche** ?

Modéliser la tâche différemment

- Séparer la tâche en plusieurs tâches plus simples
  - décomposer le problème
  - comprendre où sont les difficultés pour la tâche
- Décomposer l'information encodée par les étiquettes de relation en valeurs d'un petit ensemble de caractéristiques : les **primitives**

Approche :

- décomposition *a priori* des relations en primitives conceptuelles résultant de l'analyse (psycho-linguistique) de Sanders et collègues
- mise à l'épreuve d'une théorie / apport théorique à un modèle empirique

Cognitive Approach to Coherence Relations (CCR)

- inventaire de primitives motivées au niveau cognitif [Sanders et al. 2018]
- mappings pour PDTB, RST, SDRT
  - primitives principales [Sanders et al. 1992, 1993]
  - primitives supplémentaires : servent à expliciter les spécificités de certains cadres
- sert d'interface entre les différents cadres théoriques

Approche

- un mapping opérationnel
  - des relations annotées vers un ensemble de valeurs de primitives
  - testé sur le PDTB
- Quelles primitives sont dures à prédire ?
  - tâche de classification pour chaque primitive
- un mapping dans l'autre sens
  - d'un ensemble de valeurs de primitives vers des étiquettes de relations compatibles
  - classification de relation

Difficultés

- PDTB : hiérarchie de relation à 3 niveaux
  - ‘end-labels’ : + spécifique, niveau 2 ou 3
  - ‘intermediate labels’ : relations sous-spécifiées
- Mapping :
  - 5 primitives principales :
    - polarity, basic operation, source of coherence, implication order (2 valeurs)
    - temporal order : chronological, anti-chronological, synchronous
  - + Non-Specified : ambiguités (plusieurs valeurs possibles ou labels intermédiaires)
  - 3 primitives supplémentaires : conditional, alternative, specificity (binaire)
- Ces primitives ne sont pas de même importance dans le PDTB :
  - basic operation et polarity : distinctions entre classes de niveau 1
  - alternative, specificity : caractérise un ensemble plus restreint de relations
  - source of coherence : distinctions de niveau 3

Exemple 1

Comparison.Concession.Contra-expectation

(a) The biofuel is more expensive to produce.
(b) but by reducing the tax the government makes it possible to sell the fuel the same price.

- implication attendue : ‘the biofuel costs more’ (Q)
- présentation d’un déni de l’attente (not-Q)
  - basic operation = causal (implique une implication, sinon additive)
  - polarity = negative (implique une négation, sinon positive)
  - implication order : seulement pour relations causales (ordre prémisse / conclusion)

Relation	Basic op	Polarity	Impl. Order	SoC	Temp
Contra-expectation	causal	neg	basic	NS	NS

Exemple 2

Source of coherence : distinction commune

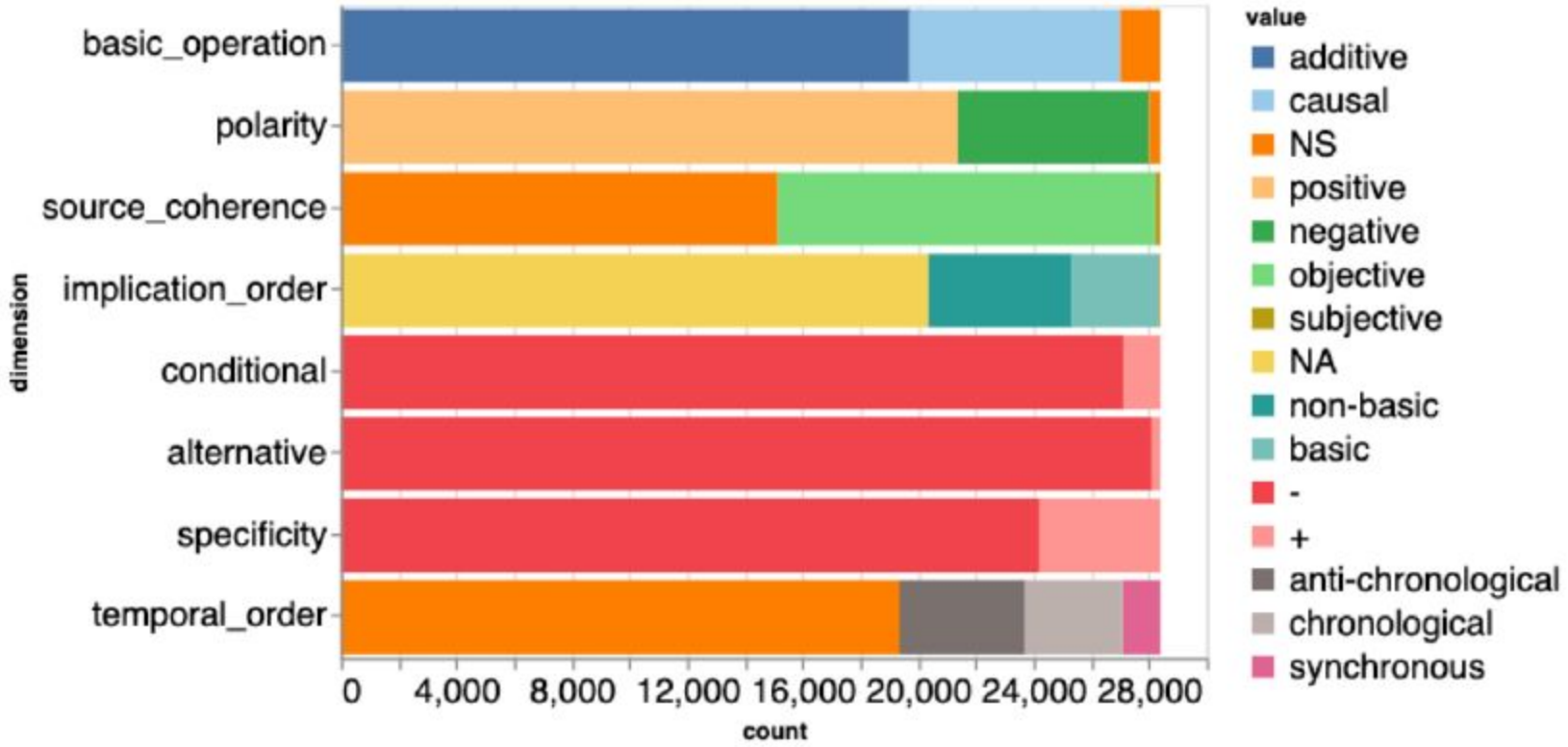
- objective : contenu propositionnel
- subjective : niveau de l'acte de parole

Contingency.Pragmatic cause.Justification

(a) Mrs Yeargin is lying.
(b) (because) They found students in an advanced class a year earlier who said she gave them similar help.

Relation	Basic op.	Polarity	Impl. Order	SoC	Temp
Justification	causal	pos	non basic	subjective	NS

Distribution de valeurs pour chaque primitive

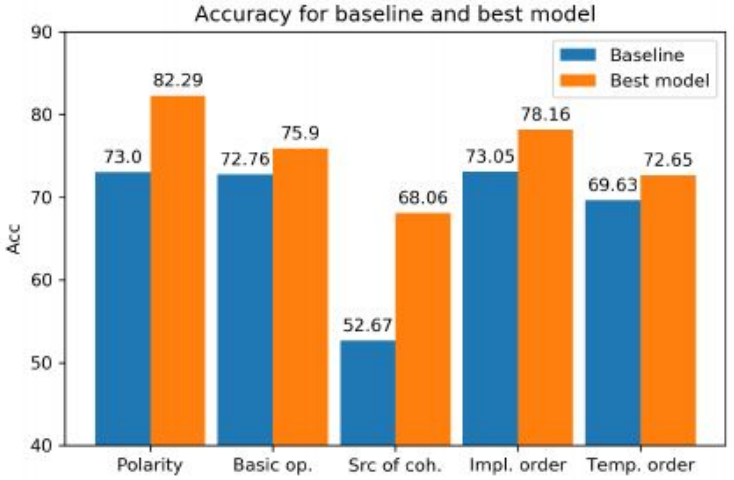


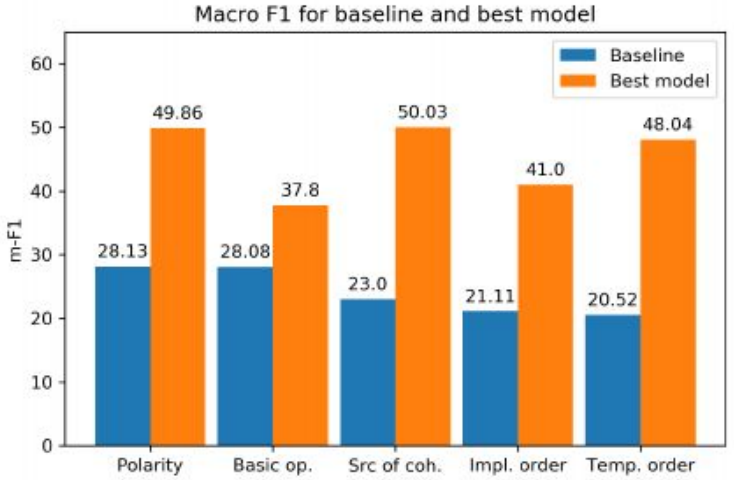
Configuration

- Classification pour chaque primitive : 28 402 paires d'arguments (train)
- Représentation de chaque argument : Infersent sentence encoder (embeddings pré-entraînés, bi-LSTM)
- Combinaison des représentations des 2 arguments :
  - concaténation
  - différence
  - produit

Résultats

- polarity et basic operation
  - primitives importantes
  - distributions similaires
- basic operation
  - amélioration la + faible
  - seulement 17% des relations causales sont classées correctement
- meilleur résultats pour polarity
  - 50% des relations négatives sont correctement classées
- source of coherence :
  - amélioration importante
  - mais moins d'1% de relations subjectives
- temporal order :
  - surtout étiquetées NS





Mapping inverse

Prédire les relations compatibles à partir des valeurs de primitives prédites

- ensemble de toutes les relations possibles (dans toute la hiérarchie)
- enlever les relations incompatibles
- enlever l'information redondante : on garde que la relation de niveau le plus haut si l'ensemble contient tous les sous-types possibles
  - Temporal, ~~Temporal.Asynchronous, Temporal.Synchrony~~

Evaluation

Plusieurs difficultés avec l'évaluation :

- sous-spécifications (étiquette de haut niveau) : mesure de classification hiérarchique
- disjonction de relations (ensemble de relations possibles) : mesure de classification multi-label

[Kiritchenko et al. 2005]

Résultats

- trop de sous-spécification
- prédiction de trop nombreux labels
- consistent avec résultats précédents : relations de type Contingency rarement prédites
  - relations Causal pour basic operation
  - souvent Temporal prédit à la place

- erreur complète alors qu'1 seule primitive est fausse

	Acc	h-R	h-P	max-h-R	max-h-P
			All		
Baseline	20.03	27.65	29.97	28.97	30.98
Primitives	34.15	28.89	19.32	49.07	**59.05**
Relations	**45.35**	**52.97**	**54.95**	**55.42**	56.58
			Explicit		
Baseline	23.5	25.35	26.13	27.02	27.33
Primitives	46.27	35.56	26.43	59.93	**69.59**
Relations	**59.08**	**63.63**	**65.3**	**67.4**	67.8
			Implicit		
Baseline	15.73	30.5	34.72	31.38	35.5
Primitives	19.12	20.63	10.52	35.61	**45.99**
Relations	**28.35**	**39.76**	**42.11**	**40.57**	42.67

Finalement :

Primitives difficiles à prédire, qui rendent la tâche difficile ?

- basic operation : l'une des + importantes très difficile à prédire

Les primitives ne sont pas indépendantes les unes des autres

- les apprendre en isolation est forcément moins précis que d'apprendre directement les étiquettes complètes
- futur : apprentissage multi-tâche

Autre extension :

- cross-corpus (RST et SDRT)
- few-shot : apprentissage sans certaines relations, peut-on les prédire quand même ?

RST Discourse parsing

(RST) Discourse Parsing

- Parsing discursif ≈ parsing syntaxique : arbres couvrant un document
  - en général, mêmes méthodes : shift-reduce, CKY, format dépendance…
- Difficultés spécifiques :
  - combinatoire / efficacité
  - représentation des données
  - manque de données (385 arbres RST)
  - + manque de rigueur dans l'évaluation, cf [Morey et al 2016]
- Beaucoup de travaux :
  - apprendre une représentation [Ji and Eisenstein,], Multi-view and multi-task training of RST discourse parsers (2016)
  - diviser la tâche : arbres nus puis classifieurs pour les relations [Wang et al, 2017]
  - études cross-lingues : Cross-lingual RST discourse parsing (2017), EusDisParser (2019)

Cross-lingual RST discourse parsing

- Petits corpus : systèmes monolingues si > 100 arbres = combiner les corpus
- Différences dans les schémas d'annotation = harmonisation

Corpus	#Arbres	#Mots	#Relations	#EDUs
EN	385	206 300	56	21 789
PT	329	135 820	32	12 573
ES	266	69 787	29	4 019
DE	173	32 274	30	2 790
NL	80	27 920	31	2 345
EU	85	27 982	31	2 396

Cross-lingual RST discourse parsing

- Petits corpus : systèmes **monolingues si > 100 arbres** = combiner les corpus
- Différences dans les schémas d'annotation = harmonisation

Corpus	#Arbres	#Mots	#Relations	#EDUs
EN	385	206 300	56	21 789
PT	329	135 820	32	12 573
ES	266	69 787	29	4 019
DE	173	32 274	30	2 790
NL	**80**	27 920	31	2 345
EU	**85**	27 982	31	2 396

Systèmes

- Monolingual systems:
  - Fully supervised, state-of-the-art performance
  - En, De, Es, Pt: > 100 trees (38 kept for test)
- Cross-lingual source only:
  - Performance when no data for the target language?
  - Train and optimize on all the source languages
- Cross-lingual source+target:
  - Improvements by combining the corpora?
  - Train on source(+target) and optimize on target

Configuration

Constituent parser [Coavoux and Crabbé, 2016]

- Lexicalized shift-reduce transition system
- Scoring system: feed-forward neural network
- Any features, mapped to real-valued vectors
- Pre-trained embeddings
- Beam search

Information from:

- the 2 EDUs on the top of the stack + EDU on the queue
- the left and right children of the 2 elements on the stack
- representation built for the 2 top elements on the stack

Traits

Features types:

- First 3 and last words + POS,
- "head set" [Sagae, 2009] vs all words [Li et al., 2014; Ji and Eisenstein, 2014]
- Position of the EDU and length, position of the head
- Number/date/percent/money

→Universal Dependencies (UDPipe)

Cross-lingual:

- Bi-lingual Wiktionaries
- Cross-lingual embeddings [Levy et al., 2017]

Résultats [Braud et al. 2017]

System	En-DT			Pt-DT			Es-DT			De-DT			Nl-DT			Eu-DT		
	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel
MFS	58.2	33.4	22.1	57.3	33.9	23.23	82.0	51.5	17.7	61.3	37.8	13.2	57.9	35.5	22.0	63.2	34.9	18.8
Li et al.	85.0	70.8	58.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ji and Eisenstein	82.1	71.1	61.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mono	85.0	72.3	60.1	82.0	65.1	49.9	89.7	72.7	54.4	80.2	53.9	35.0	-	-	-	-	-	-
+ emb.	83.5	68.5	55.9	81.3	62.9	48.8	89.3	72.4	51.4	77.7	51.6	31.1	-	-	-	-	-	-
Source only	76.3	50.5	31.3	76.5	54.6	35.5	78.1	45.4	27.0	76.0	46.0	26.1	69.5	42.1	25.3	78.6	53.0	26.4
Source+Target	85.1	73.1	61.4	81.9	65.1	49.8	88.8	68.0	50.4	79.6	53.6	34.1	69.2	43.4	28.3	76.7	50.5	29.5

Résultats [Braud et al. 2017]

System	En-DT			Pt-DT			Es-DT			De-DT			Nl-DT			Eu-DT		
	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel
MFS	58.2	33.4	22.1	57.3	33.9	23.23	82.0	51.5	17.7	61.3	37.8	13.2	57.9	35.5	22.0	63.2	34.9	18.8
Li et al.	85.0	70.8	58.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ji and Eisenstein	82.1	71.1	61.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mono	85.0	72.3	60.1	82.0	65.1	49.9	89.7	72.7	54.4	80.2	53.9	35.0	-	-	-	-	-	-
+ emb.	83.5	68.5	55.9	81.3	62.9	48.8	89.3	72.4	51.4	77.7	51.6	31.1	-	-	-	-	-	-
Source only	76.3	50.5	31.3	76.5	54.6	35.5	78.1	45.4	27.0	76.0	46.0	26.1	69.5	42.1	25.3	78.6	53.0	26.4
Source+Target	85.1	73.1	61.4	81.9	65.1	49.8	88.8	68.0	50.4	79.6	53.6	34.1	69.2	43.4	28.3	76.7	50.5	29.5

Résultats [Braud et al. 2017]

System	En-DT			Pt-DT			Es-DT			De-DT			Nl-DT			Eu-DT		
	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel
MFS	58.2	33.4	22.1	57.3	33.9	23.23	82.0	51.5	17.7	61.3	37.8	13.2	57.9	35.5	22.0	63.2	34.9	18.8
Li et al.	85.0	70.8	58.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ji and Eisenstein	82.1	71.1	61.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mono	85.0	72.3	60.1	82.0	65.1	49.9	89.7	72.7	54.4	80.2	53.9	35.0	-	-	-	-	-	-
+ emb.	83.5	68.5	55.9	81.3	62.9	48.8	89.3	72.4	51.4	77.7	51.6	31.1	-	-	-	-	-	-
Source only	76.3	50.5	31.3	76.5	54.6	35.5	78.1	45.4	27.0	76.0	46.0	26.1	69.5	42.1	25.3	78.6	53.0	26.4
Source+Target	85.1	73.1	61.4	81.9	65.1	49.8	88.8	68.0	50.4	79.6	53.6	34.1	69.2	43.4	28.3	76.7	50.5	29.5

Source only = e.g.:

- train on En+Pt+Es+De
- test on Eu

Résultats [Braud et al. 2017]

System	En-DT			Pt-DT			Es-DT			De-DT			Nl-DT			Eu-DT		
	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel
MFS	58.2	33.4	22.1	57.3	33.9	23.23	82.0	51.5	17.7	61.3	37.8	13.2	57.9	35.5	22.0	63.2	34.9	18.8
Li et al.	85.0	70.8	58.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ji and Eisenstein	82.1	71.1	61.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mono	85.0	72.3	60.1	82.0	65.1	49.9	89.7	72.7	54.4	80.2	53.9	35.0	-	-	-	-	-	-
+ emb.	83.5	68.5	55.9	81.3	62.9	48.8	89.3	72.4	51.4	77.7	51.6	31.1	-	-	-	-	-	-
Source only	76.3	50.5	31.3	76.5	54.6	35.5	78.1	45.4	27.0	76.0	46.0	26.1	69.5	42.1	25.3	78.6	53.0	26.4
Source+Target	85.1	73.1	61.4	81.9	65.1	49.8	88.8	68.0	50.4	79.6	53.6	34.1	69.2	43.4	28.3	76.7	50.5	29.5

Source only = e.g.:

- train on En+Pt+Es+De
- test on Eu

Résultats [Braud et al. 2017]

System	En-DT			Pt-DT			Es-DT			De-DT			Nl-DT			Eu-DT		
	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel
MFS	58.2	33.4	22.1	57.3	33.9	23.23	82.0	51.5	17.7	61.3	37.8	13.2	57.9	35.5	22.0	63.2	34.9	18.8
Li et al.	85.0	70.8	58.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ji and Eisenstein	82.1	71.1	61.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mono	85.0	72.3	60.1	82.0	65.1	49.9	89.7	72.7	54.4	80.2	53.9	35.0	-	-	-	-	-	-
+ emb.	83.5	68.5	55.9	81.3	62.9	48.8	89.3	72.4	51.4	77.7	51.6	31.1	-	-	-	-	-	-
Source only	76.3	50.5	31.3	76.5	54.6	35.5	78.1	45.4	27.0	76.0	46.0	26.1	69.5	42.1	25.3	78.6	53.0	26.4
Source+Target	85.1	73.1	61.4	81.9	65.1	49.8	88.8	68.0	50.4	79.6	53.6	34.1	69.2	43.4	28.3	76.7	50.5	29.5

Source + Target = train on Pt+Es+De+En ; test on En

- improvement for relations largely represented in all corpora (e.g. JOINT +3%),
- or under-represented in the EnDT (e.g. CONDITION +3%).

Matrice de confusion (Basque)

RST relation	Match	
ELABORATION	101	0.616
SAME-UNIT	40	0.244
EVALUATION	9	0.055
BACKGROUND	6	0.036
MEANS	3	0.018
CAUSE	2	0.012
ENABLEMENT	2	0.012
JOINT	1	0.006
Total agreement	164	

Table 7: Description of gold and automatic label matching

Relation	Errors	Empl. Tags
ELABORATION	213	314
BACKGROUND	252	258
JOINT	191	192
CAUSE	77	79
SAME-UNIT	67	107
EVALUATION	63	72
ENABLEMENT	29	31

Table 8: Parser annotation confusion matrix

- machine tries to get the best results using a small number of very general relations such as: ELABORATION, BACKGROUND and JOINT
- disagreement between humans : general, widely used and less informative relations, such as ELABORATION, LIST, BACKGROUND, RESULT and MEANS

Finalement :

- difficile de combiner des corpus : problème de langues / adaptation ? ou de différences dans les schémas d'annotations ?
- + problème d'évaluation : un système qui prédit surtout bien des élaborations n'est peut-être pas très utile ...

Multi-task and multi-view training of RST parsers

- Motivation: lack of data and large range of information needed
- Solution: leveraging knowledge from other data

→**Multi-task learning**

Previously used for discourse relation identification [Lan et al., 2013, Liu and Li, 2016]

Combiner plusieurs corpus

- **Discourse corpora** following different frameworks: RST DT and PDTB
- Data for **other related tasks** (sentences) : temporality, factuality, coreference, speech turn...
- “**Multi-view**": different representations of the output, i.e. dependencies, label granularity

Setting :

- Neural network (bi-lstm), sequence prediction

- Hard parameter sharing: sharing internal layers

Les ‘vues’ et les ‘tâches’

Alternate views

	Constituent (main)	Dependency	Nuclearity	Relation	Fine-grained
EDU 1	(NN-TextOrg(NN-SaUnit(NN-List	Root	(NN(NN(NN	(TextOrg(SaUnit(List	(TextOrg(SaUnit(List
EDU 2	NN-List)	-1 NN-List	NN)	List)	List)
EDU 3	(NS-Elab	-2 NN-SaUnit	(NS	(Elab	(Elab-set-member

Auxiliary tasks

	Speech	Factuality	Aspect	Modality	Polarity	Tense	Coreference	PDTB
Corpus	Santa Barbara	Factbank	Timebank	Timebank	Timebank	Timebank	Ontonotes	PDTB
Sent 1	turn1	Certain	Progressive	Must	Positive	Past	Root	Root
Sent 2	turn2	Probable	Perfective	Could	Negative	Future	Coreferent	Contrast

Résultats

System	Fine	Fact	Speech	Asp	Dep	Nuc+lab	Mod	Pol	PDTB	Coref	Ten	Span	Nuclearity	Relation
Prior work														
DPLP concat	-	-	-	-	-	-	-	-	-	-	-	82.08	**71.13**	61.63
DPLP general	-	-	-	-	-	-	-	-	-	-	-	81.60	70.95	**61.75**
Feng and Hirst [2014]	-	-	-	-	-	-	-	-	-	-	-	84.9	69.9	57.2
Feng and Hirst [2014] PE	-	-	-	-	-	-	-	-	-	-	-	**85.7**	71.0	58.2
Our work														
Hier-LSTM	-	-	-	-	-	-	-	-	-	-	-	81.39	64.54	49.15
MTL-Hier-LSTM	✓	-	-	-	-	-	-	-	-	-	-	82.88	67.46	53.25
MTL-Hier-LSTM	-	✓	-	-	-	-	-	-	-	-	-	83.40	67.16	52.10
MTL-Hier-LSTM	-	-	✓	-	-	-	-	-	-	-	-	83.26	67.51	51.75
MTL-Hier-LSTM	-	-	-	✓	-	-	-	-	-	-	-	83.69	66.25	51.25
MTL-Hier-LSTM	-	-	-	-	✓	-	-	-	-	-	-	81.25	65.34	51.24
MTL-Hier-LSTM	-	-	-	-	-	✓	-	-	-	-	-	82.09	65.68	51.12
MTL-Hier-LSTM	-	-	-	-	-	-	✓	-	-	-	-	81.66	65.31	50.58
MTL-Hier-LSTM	-	-	-	-	-	-	-	✓	-	-	-	82.01	65.29	50.11
MTL-Hier-LSTM	-	-	-	-	-	-	-	-	✓	-	-	81.61	63.10	48.89
MTL-Hier-LSTM	-	-	-	-	-	-	-	-	-	✓	-	80.26	63.35	47.70
MTL-Hier-LSTM	-	-	-	-	-	-	-	-	-	-	✓	81.33	62.34	47.57
Best combination	-	-	-	-	✓	✓	✓	-	✓	-	-	83.62	69.77	55.11
Human annotation	-	-	-	-	-	-	-	-	-	-	-	88.70	77.72	65.75

- MTL improves over STL for 8/11tasks
- Alternate views are the most beneficial
- Fact and Speech are the most beneficial auxiliary tasks
- Best system: Task combinations
- Tense and Coreference: No improvement, calls for a finer grained encoding

Conclusion

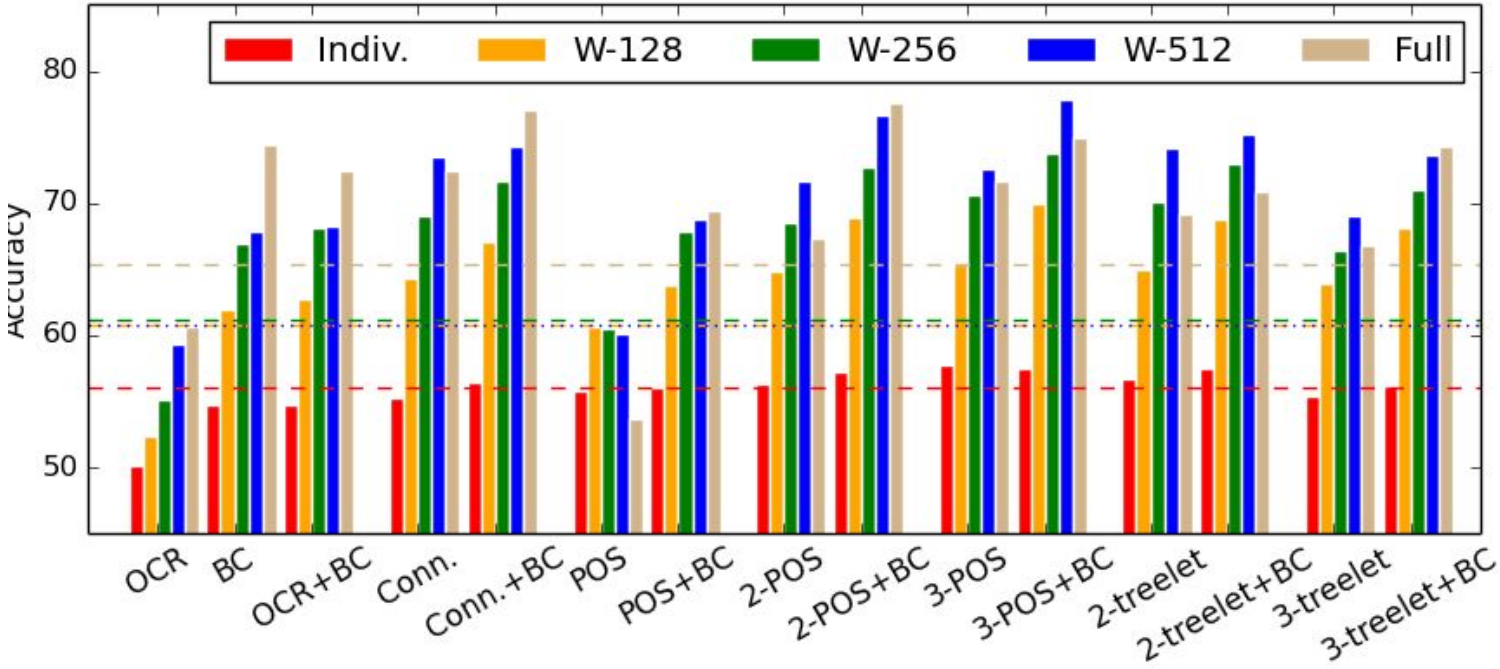
- L’évaluation
- Les défis actuels

Evaluation applicative

- Scores d'évaluation ne veulent pas dire grand chose, cf EusDisParser ou [Morey et al. 2016]
- Quelles informations discursives sont utiles ?
  - a-t-on besoin des 56 relations RST ?
  - a-t-on besoin du span exact des arguments PDTB ?
  - a-t-on besoin d'arbres couvrant ou chunking suffit ?
- Travaux sur l'analyse de sentiment [Bahtia et al. 2015] : mais exagéré ?
- Réflexion en cours ...
  - Détection de fraude scientifique avec relations explicites [Braud and Søgaard, 2016]
  - Comparaison discours / argumentation [Huber et al. 2019]
  - Parole altérée : langage des schizophrènes, modélisation du dialogue (thèse en cours)

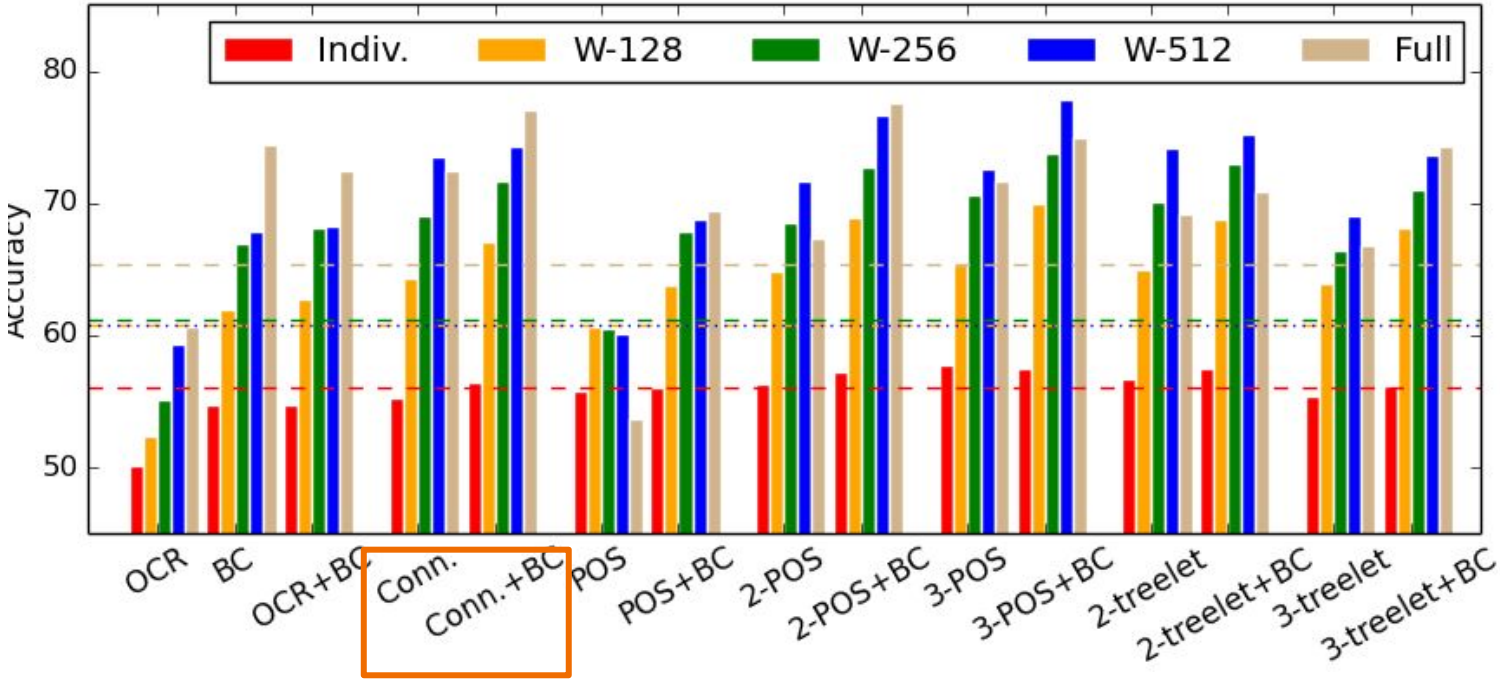
Evaluation applicative

Identification des marqueurs langagiers de la schizophrénie



Evaluation applicative

Identification des marqueurs langagiers de la schizophrénie



Défis actuels

- Toujours gérer le manque de données (annotation très coûteuse)
  - [Badene et al 2019] : annotation distante
  - [Carrenini et al 2019] : utiliser la tâche de détection de sentiment pour inférer la structure discursive d'un document (distant learning), résultat mitigé (structure + nucléarité)
  - [Liu and Lapata, 2017; Karimi et al, 2019] : structures latentes
    - Etude des biais dans les articles de journaux (thèse en cours)
- Note : on a quand même besoin d'annotations :)
  - idéalement en variant langues / domaines / modalités
- Toujours sortir de l'anglais, du WSJ et des monologues
  - Nouvelle shared task DisRPT (CODI 2021)
- Nécessité d'une réflexion linguistique
  - Différences entre formalismes : comment les concilier ?
  - Différences entre les langues / domaines / modalités ?
  - Quelles informations sont nécessaires pour l'analyse discursive ? comment les intégrer ?
  - Comment utiliser l'information discursive pour des applications ?

Champ de recherche en plein essor

- De plus en plus de publications dans les conférences, surtout :
  - identification des relations implicites
  - parsing RST
  - léger regain pour la segmentation
- Workshops dédiés :
  - LSDsem, DiscoMT, DisRPT, DSSNLG
  - CoDi in 2020 and 2021 !
  - (and maybe a SIG :)
- Intérêt industriel :
  - quantum (anr) : prise en compte de la structure des documents pour la génération de question (amélioration de FAQs)

Références (+ pub Melodi)

- Données de segmentation https://github.com/disrpt/sharedtask2019/tree/master/data
- Segmenteur “utilisable” pour le Français: bientôt en ligne/me contacter
- Décomposition des relations du PDTB (conditionnée aux droits sur le PTB): nous contacter

Le discours est une thématique importante dans l’équipe Melodi:

- Sileo D, Van De Cruys T, Pradel C and Muller P, *Mining Discourse Markers for Unsupervised Sentence Representation Learning*, (NAACL 2019)
- Morey, M., Muller, P., and Asher, N. *A dependency perspective on RST discourse parsing and evaluation*. Computational Linguistics (2018).
- Mathieu Morey, Philippe Muller, Nicholas Asher. *How much progress have we made on RST discourse parsing? A replication study of recent results on the RST-DT* (EMNLP 2018, short paper)
- Sonia Badene, Kate Thompson, Jean-Pierre Lorré and Nicholas Asher, *Weak Supervision for Learning Discourse Structure.* (EMNLP 2019)

Merci !

*And even in our wildest and most wandering reveries, nay in our very dreams, we shall find, if we reflect, that **the imagination ran not altogether at adventures**, but that there was still a connection upheld among the different ideas, which succeeded each other. Were the loosest and freest conversation to be transcribed, there would immediately be transcribed, there would immediately be observed **something which connected it in all its transitions**.*

David Hume, An enquiry concerning human understanding, 1748