Analyse discursive: segmentation automatique et prédiction de relations rhétoriques

Philippe Muller, Chloé Braud, Charlotte Roze, Mathieu Morey

Equipe Melodi,
IRIT, Univ. Toulouse

Equipe Synalp
LORIA/CNRS Nancy

Datactivist
Aix en Provence

CLLE 30/9/2019

Vue d’ensemble

- Deux expériences d’analyse automatique du discours
  - segmentation automatique de textes en unités discursives (multilingue)
  - prédiction de relations de discours entre unités discursives données (anglais)

- Deux visions différentes de l’apport de la linguistique au TAL
  - segmentation: en entrée, texte tokenisé et utilisation d’embeddings contextuels
    - “pas de linguistique” (?)
  - relations de discours: 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

Plan

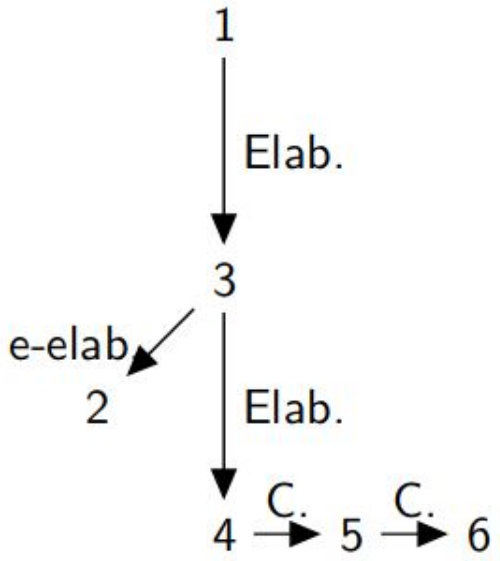
- Introduction (très) rapide à l’analyse du discours
- Segmentation automatique multilingue
- Décomposition de relations de discours

En bonus éventuel, discussion plus précise des détails techniques

Analyse discursive

[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

[Principles of natural selection.]_1 [The theory of natural selection, [as it was initially described by Charles Darwin]_2, lies upon three principles:]_3 [1. the principle of variation]_4 [2. the principle of adaptation]_5 [3. the principle of heredity]_6



Analyse discursive : quelques hypothèses

- un texte (ou une conversation) est composé d’unités (propositions, phrases, énoncés, tour de parole)
- ces unités forment une structure, et leurs liens guident l’interprétation
  - anaphores
  - implicatures
  - ordre temporel

Exemples de relations

Explicitement marqué (“marqueurs de discours”)

(1) Climate change is caused by anthropic activities, but politics are not doing anything about it.

-> Concession

(2) Climate is changing. Humans generate too much CO_2 .

-> Reason

Exemple “local”, mais les relations peuvent être à plus longue distance.

Divers cadres théoriques

Plusieurs cadres (PDTB, RST, SDRT, ...) mais pas de consensus:

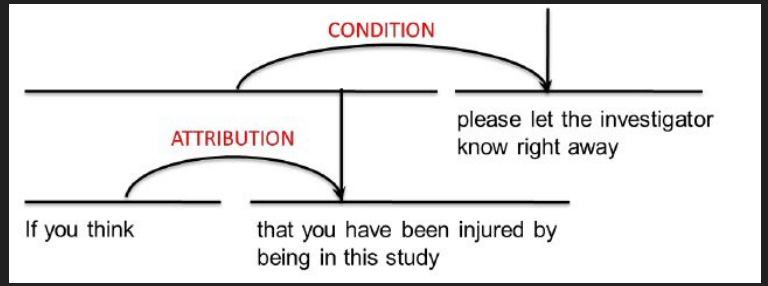
- sur la nature des unités (notamment à l'intérieur de la phrase)
- unités simples, complexes, enchassées ou pas
- sur la nature de la structure discursive (arbre, graphe, etc)
- sur les types de relations entre entités
- sur la sémantique de ces relations
- plusieurs corpus annotés, parfois dans le même cadre et la même langue, avec des directives différentes

Analyse automatique

Plusieurs étapes généralement considérées:

1. la segmentation en unités
    - éventuellement le regroupement en unités complexes
    - éventuellement repérage de marqueurs explicites de relations
2. la détermination des liens entre unités (attachement; prédiction structurelle)
3. la caractérisation du lien (étiquetage)

Ici on ne parlera que de 1 et 3



Segmentation automatique

- le début de la chaine … donc cruciale
- et pourtant très négligée: supposée “facile” donc presque tous les travaux en analyse discursive supposent une segmentation parfaite donnée
- mais peu d’études en dehors de quelques langues
- pas de comparaison entre langues, cadres théoriques …

Pour combler ce vide: tâche commune (shared task) à l’atelier Disrpt 2019

“Discourse Relation Parsing and Treebanking”

Des données !

15 corpus

10 langues

styles divers: oral transcrit, écrit, chat, …

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

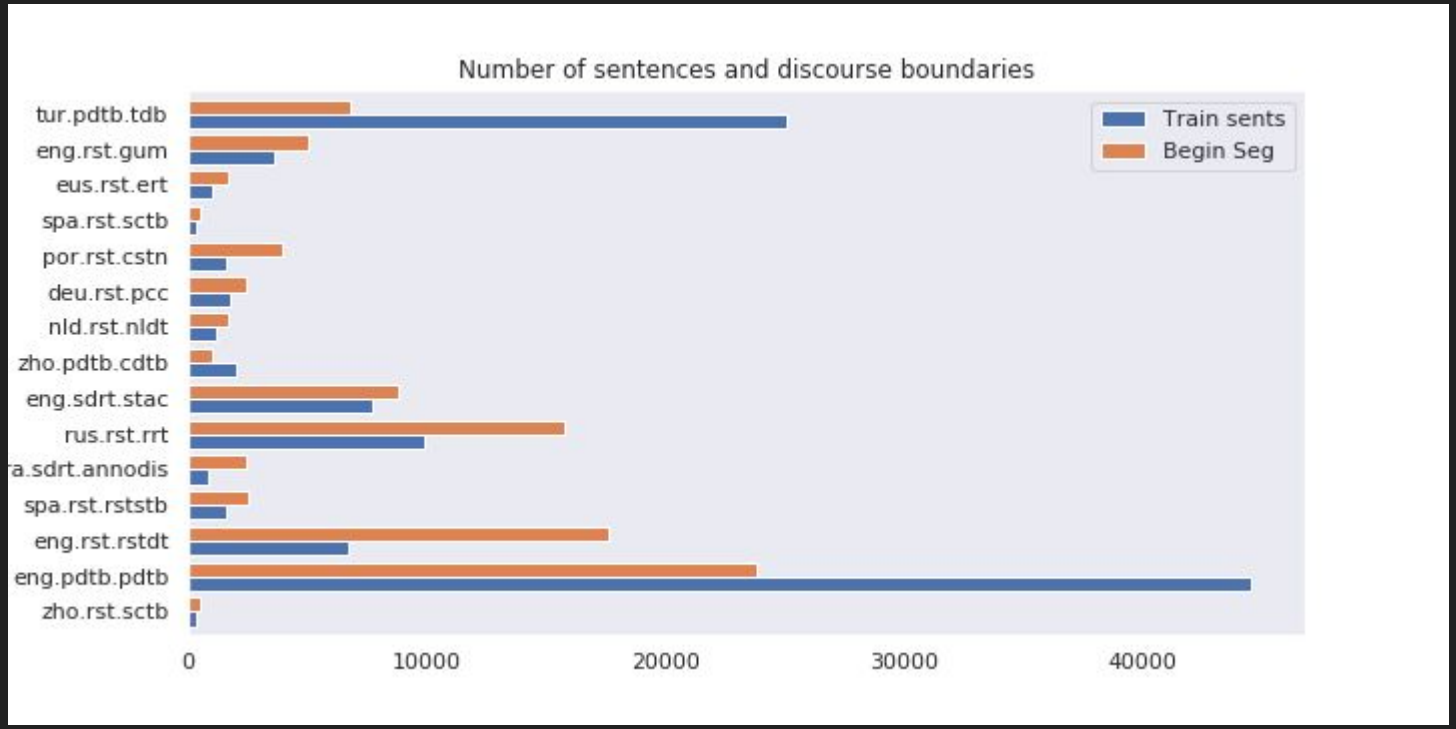
Un format unique

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

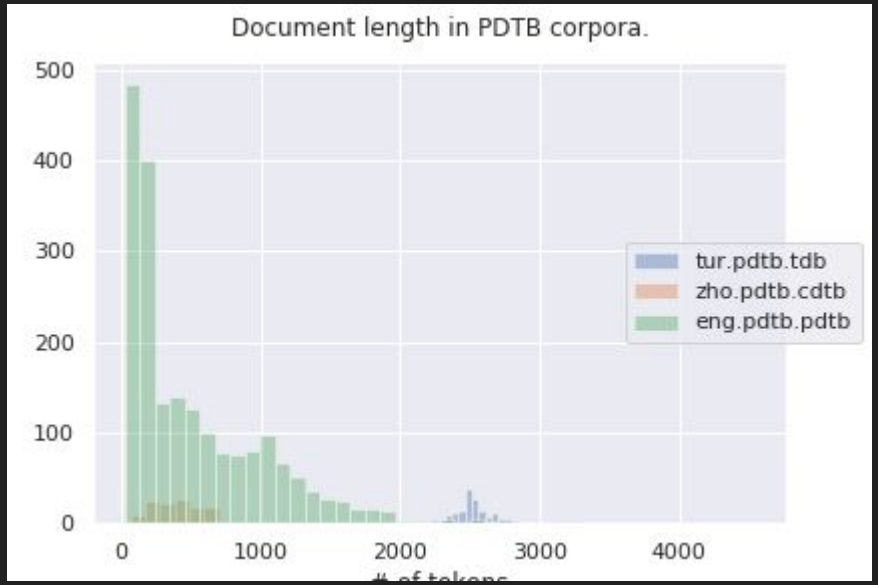
- format begin-inside-outside … ou presque
- permet une analyse en séquence directement
- deux versions: juste tokenisé ou (découpage en phrase donné + analyse syntaxique automatique)
- a du tordre un peu le bras de certains formalismes (SDRT)

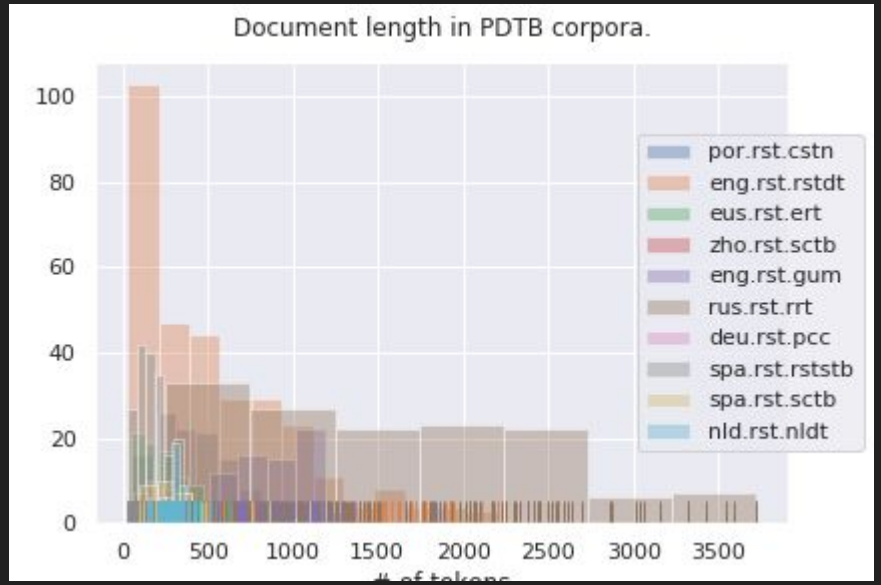
President	BeginSeg=Yes
Bush	_
insists	_
it	BeginSeg=Yes
would	_
be	_
a	_
great	_
tool	_
for	BeginSeg=Yes
curbing	_
the	_
budget	_
deficit	_
and	BeginSeg=Yes
slicing	_
(...)	

Quelques statistiques: nb de phrases & segments



Quelques statistiques: longueurs des phrases





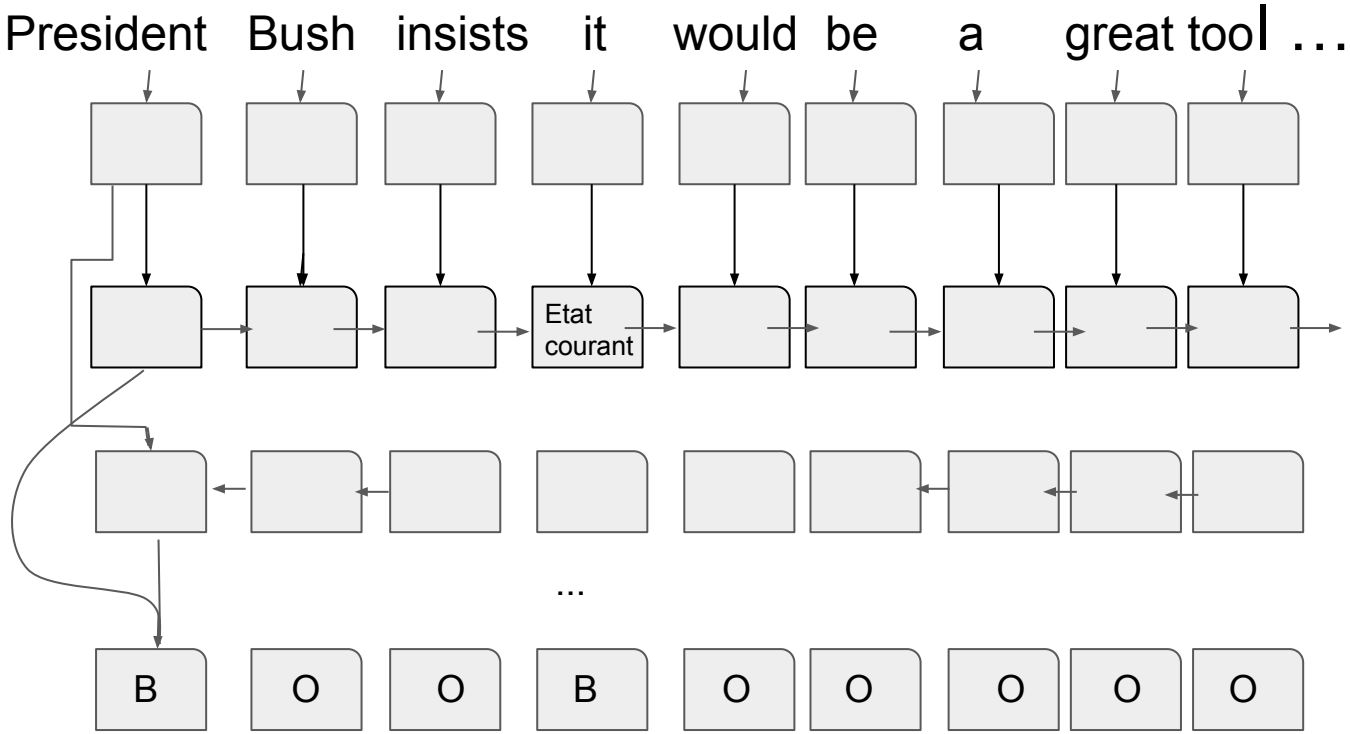
Quelques questions “scientifiques” permises par la tâche

- 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 inter formalismes / intra-langues ?
- comparaisons intra formalismes / intra-langues ?

Notre expérimentation

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 / 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



Embedding: Caractères+ Mot

LSTM G->D

LSTM D->G

Prédictions

Résultats préliminaires

Comparaison d'approches sur des 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**
RSTDT	78.97	83.21	94.41	93.72	**94.75**
STAC	77.43	71.70	84.65	84.45	**86.06**

En pratique

- 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 quand même (russe, turc)
    - prétraitements spécifiques
- Impossible de contrôler chaque langue vu les délais: souci possible pour le mandarin.

Résultats de la campagne (Disrpt 2019)

Peu de soumission (4 équipes), mais approches différentes

suspense …

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

Résultats “internes”

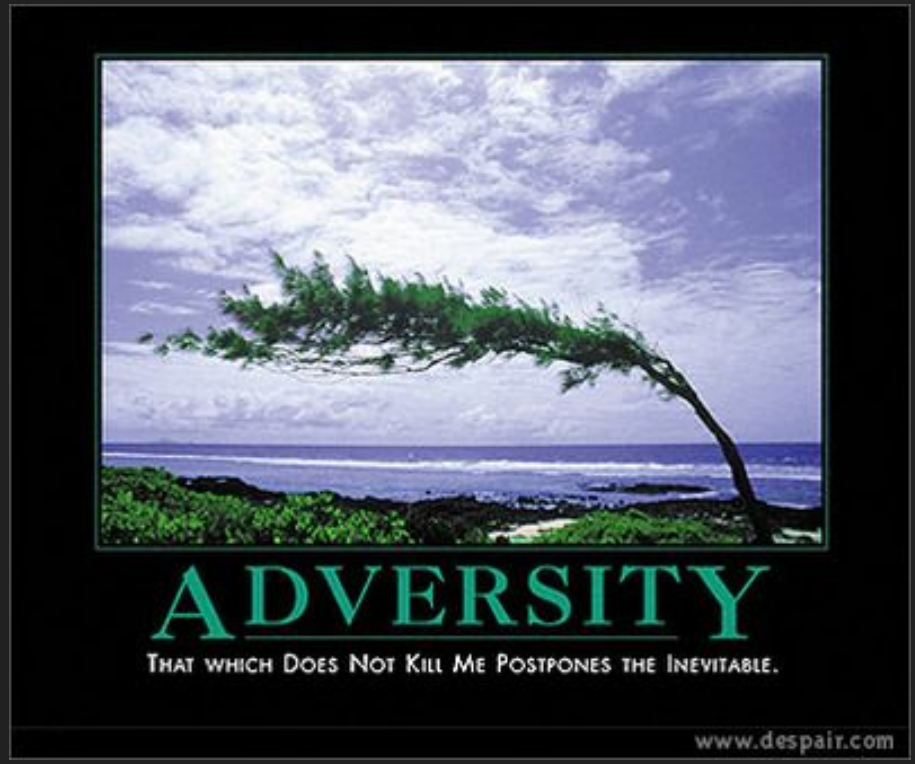
Comparaison de transfert entre corpus d’une même langue avec le même formalisme (RST/Anglais)

- RSTDT = articles de journeaux
- GUM = genres mélangés (news, académique, opinion, voyage, interviews, bio, fiction)

Train/Test	RSTDT	GUM
English RSTDT	93	73
English GUM	66	96

Quelles leçons ?

- le pouvoir des embeddings contextuels
- questionnement sur les fondements des corpus discursifs
- analyse post-hoc des erreurs encore à faire



Partie II: prédiction de relations par décomposition

Inverse de la précédente:

- une tâche difficile qui intéresse beaucoup de monde
- des questions linguistiques
- une expérimentation relativement décevante (pour l’instant !)

Et je passe la parole à

Bilan:

- Plein de choses à faire sur le discours en TAL, avec une place pour une réflexion sur les phénomènes linguistiques
- Le “boom” sémantique récent fait glisser vers la prise en compte de plus de phénomènes pragmatiques (cf la foison de “benchmarks”)
- Les nouvelles techniques et modèles à la mode peuvent être intégrées à des approches spécifiques relativement facilement grâce à la modularité des outils disponibles

Ressources

- 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

Références

- Muller P, Braud C and Morey M , ToNy: *Contextual embeddings for accurate multilingual discourse segmentation of full documents*, Workshop on Discourse Relation Parsing and Treebanking 2019. Minneapolis, pp. 115-124. Association for Computational Linguistics.
- Roze C, Braud C and Muller P. *Which aspects of discourse relations are hard to learn? Primitive decomposition for discourse relation classification*. Sigdial 2019.

Pub éhontée (suite)

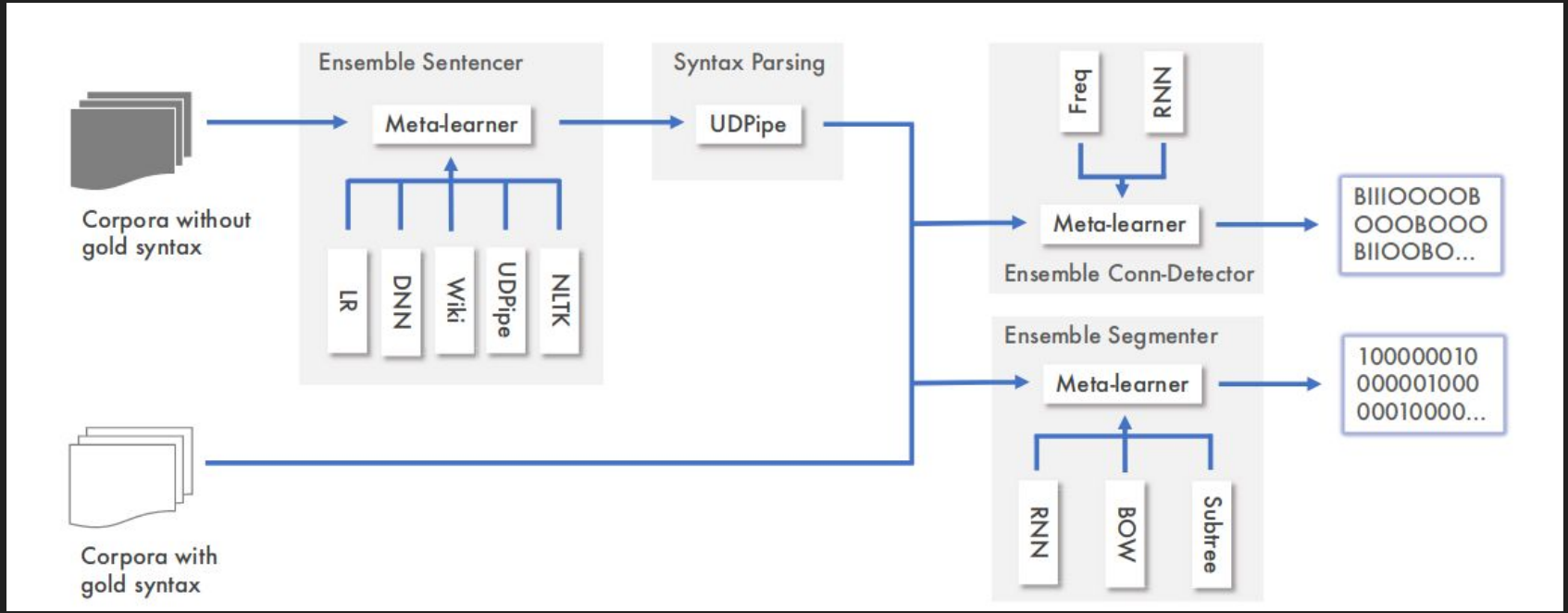
Le discours est une thématique importante dans l'équipe Melodi:

- 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)

Questions ?

Bonus track

Segmentation: 2e système de la campagne

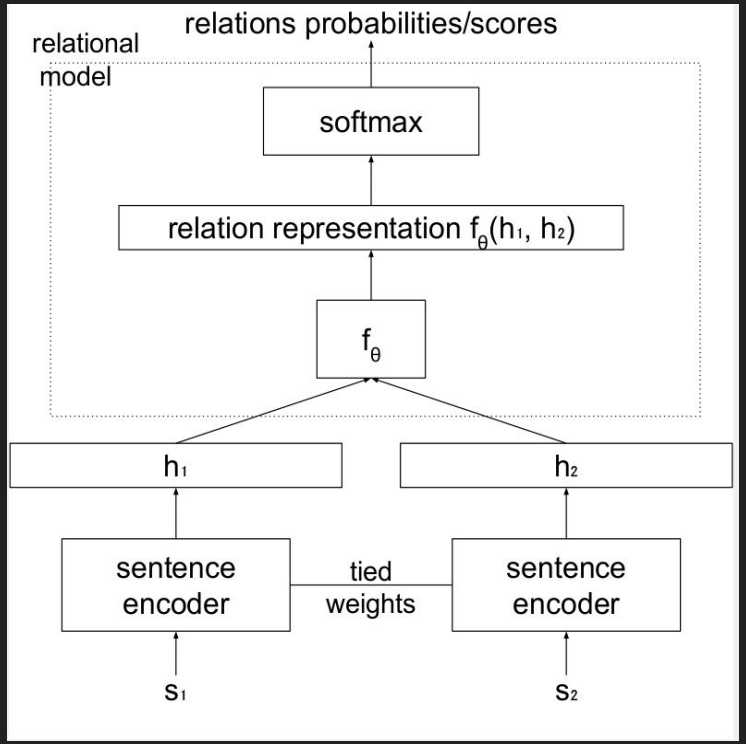


Generic relational learning supervised architecture

- Relation prediction based on composition f of two sentences representations (optionally with subsequent layers)
- Each sentence is encoded keeping the an intermediate representation, a popular choice being :
  - encoder = LSTM
  - states h_i = final state of LSTM
- Input : pretrained word embeddings for each sentence

Infersent system

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, Antoine Bordes: *Supervised Learning of Universal Sentence Representations from Natural Language Inference Data*. EMNLP 2017



Contextual embeddings = Language models

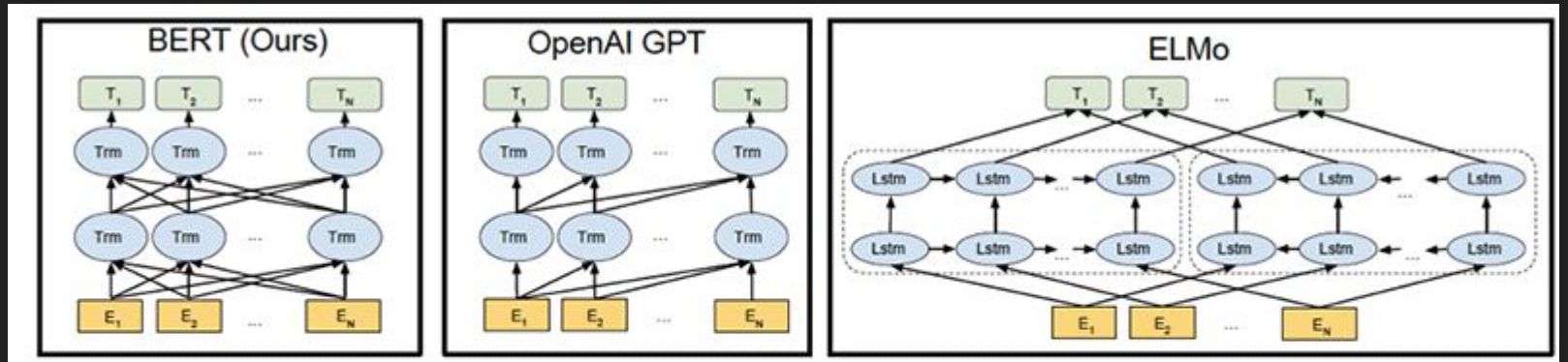
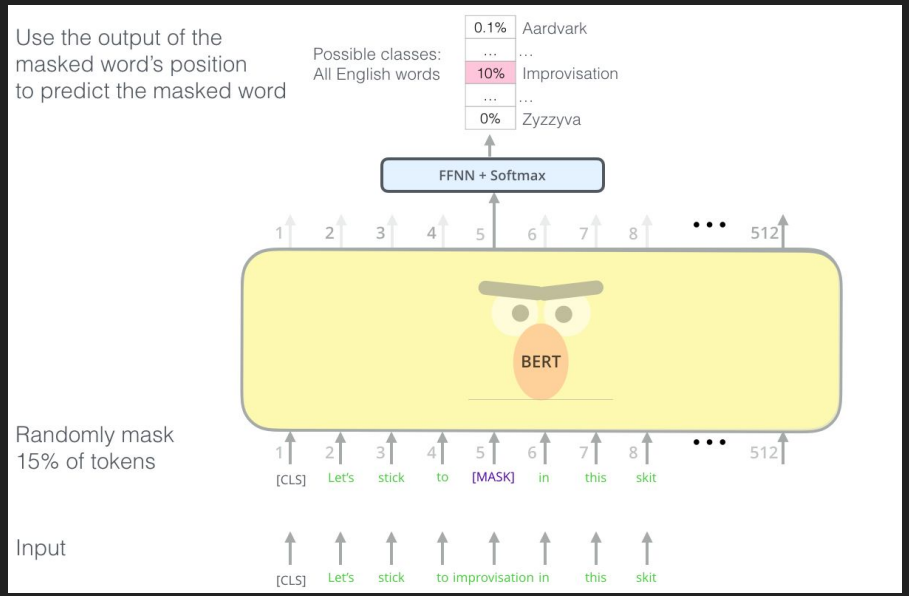
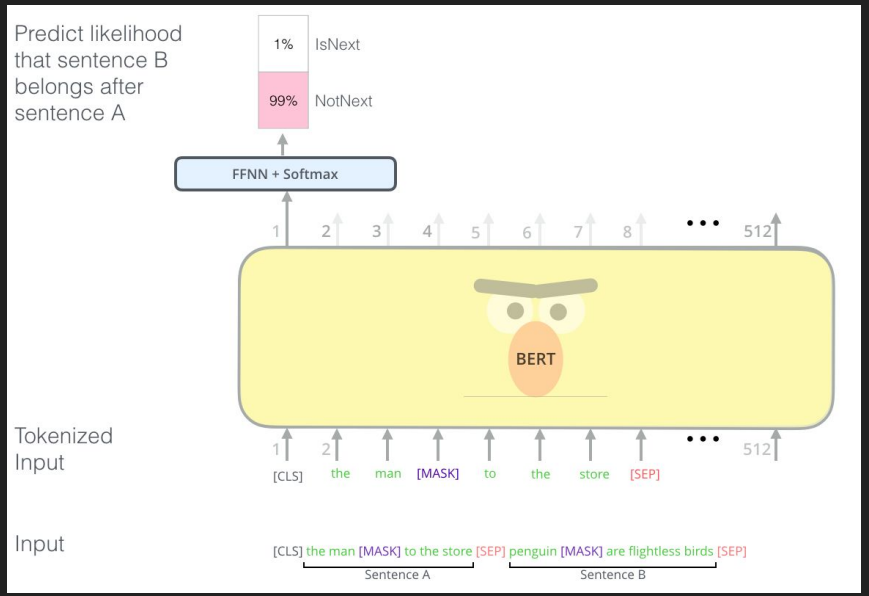


Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks. Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

Contextual embeddings: BERT

Intègre aussi un aspect inter-phrastique





Transfert et “fine-tuning”

Une fois entrainé comme un LM, BERT peut-être ajusté à une tâche spécifique ayant moins de données:

Exemple en classification de relations.