



Langage, IA générative et robotique

ANITI Days 2024

25 novembre 2024

ANITI

Apprentissage low-resource : le cas des structures discursives

Philippe Muller & Chloé Braud

November 25, 2024




Apprentissage low-resource : le cas des structures discursives - P. Muller et C. Braud

- (1) Document-level NLP: beyond sentence boundaries
- (2) What is discourse structure?
- (3) Main issues with analyzing discourse automatically
- (4) Proposed approaches for low-resource settings, multilinguality and transfer

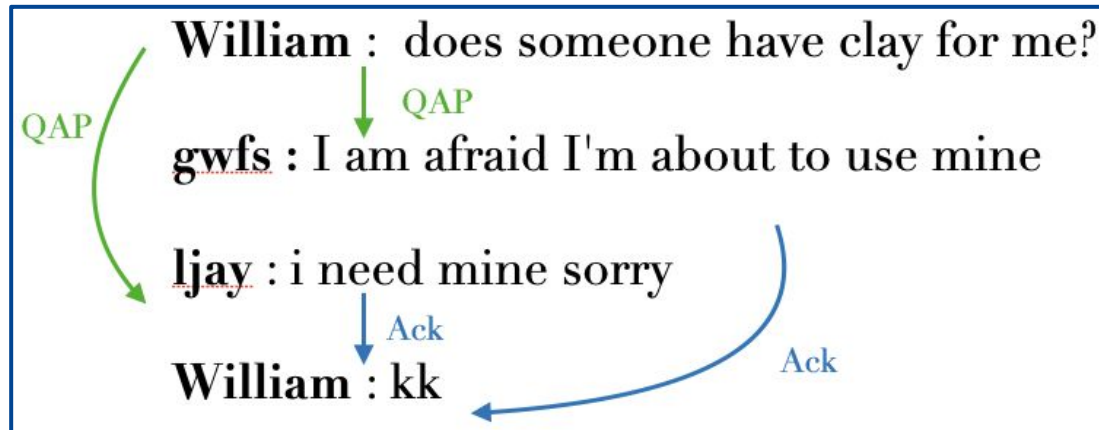
ANITI Document-level NLP: example

Recurrent models typically factor computation along the symbol positions of the input and output sequences. Aligning the positions to steps in computation time, they generate a sequence of hidden states h_t , as a function of the previous hidden state h_{t-1} and the input for position t . This inherently sequential nature precludes parallelization within training examples, which becomes critical at longer sequence lengths, as memory constraints limit batching across examples. Recent work has achieved significant improvements in computational efficiency through factorization tricks [18] and conditional computation [26], while also improving model performance in case of the latter. The fundamental constraint of sequential computation, however, remains.



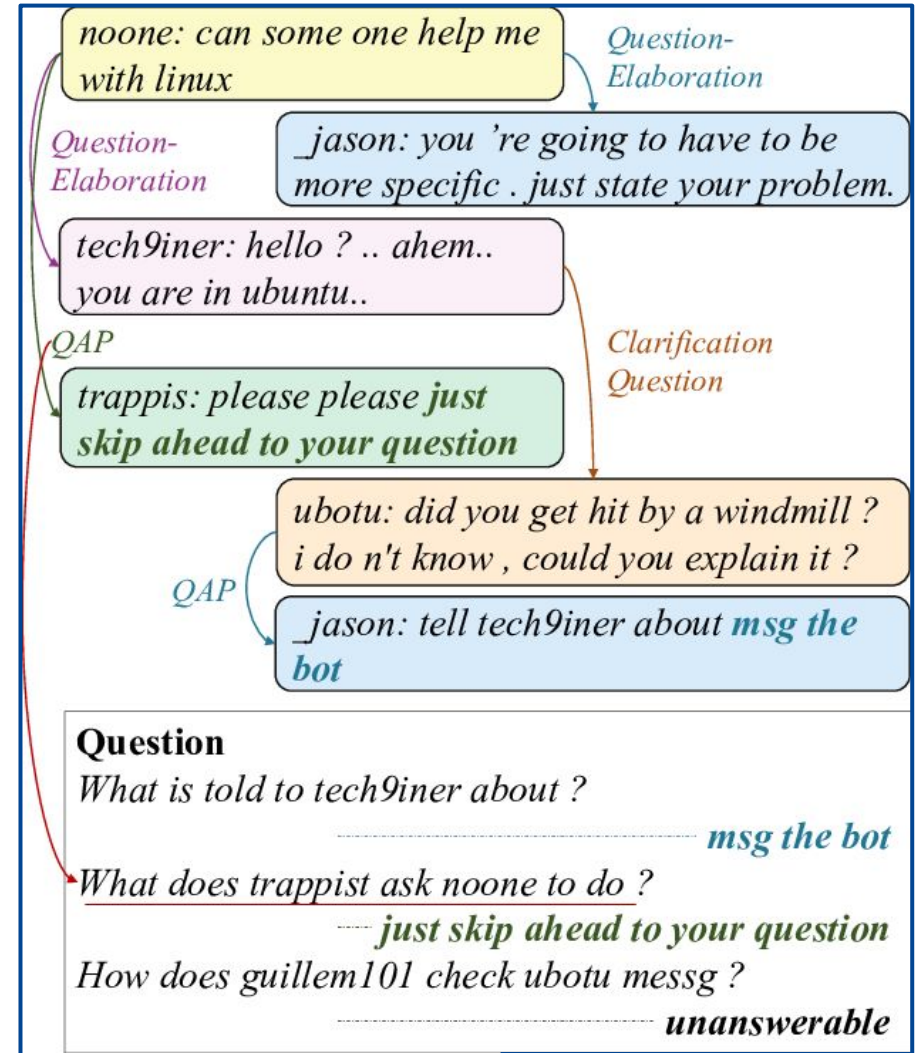
Non-local phenomena:

- anaphora
- abstract references
- textual coherence / argumentation
- mostly implicit, with some explicit clues



STAC

- relations between text units: encode the coherence
 - e.g. Question-Answer Pair, Acknowledgment, Clarification, Explanation, Result, Contrast ...
- structure: tree / graph over the document



Question
 What is told to tech9iner about ?
 msg the bot
 What does trappist ask noone to do ?
 just skip ahead to your question
 How does guillem101 check ubotu messg ?
 unanswerable

Molweni Ubuntu

ANITI Textual coherence and discourse/dialogue structure

What is at stake:

- implicit information
- intentions of the writer / of dialogue participants
- dialogue threads

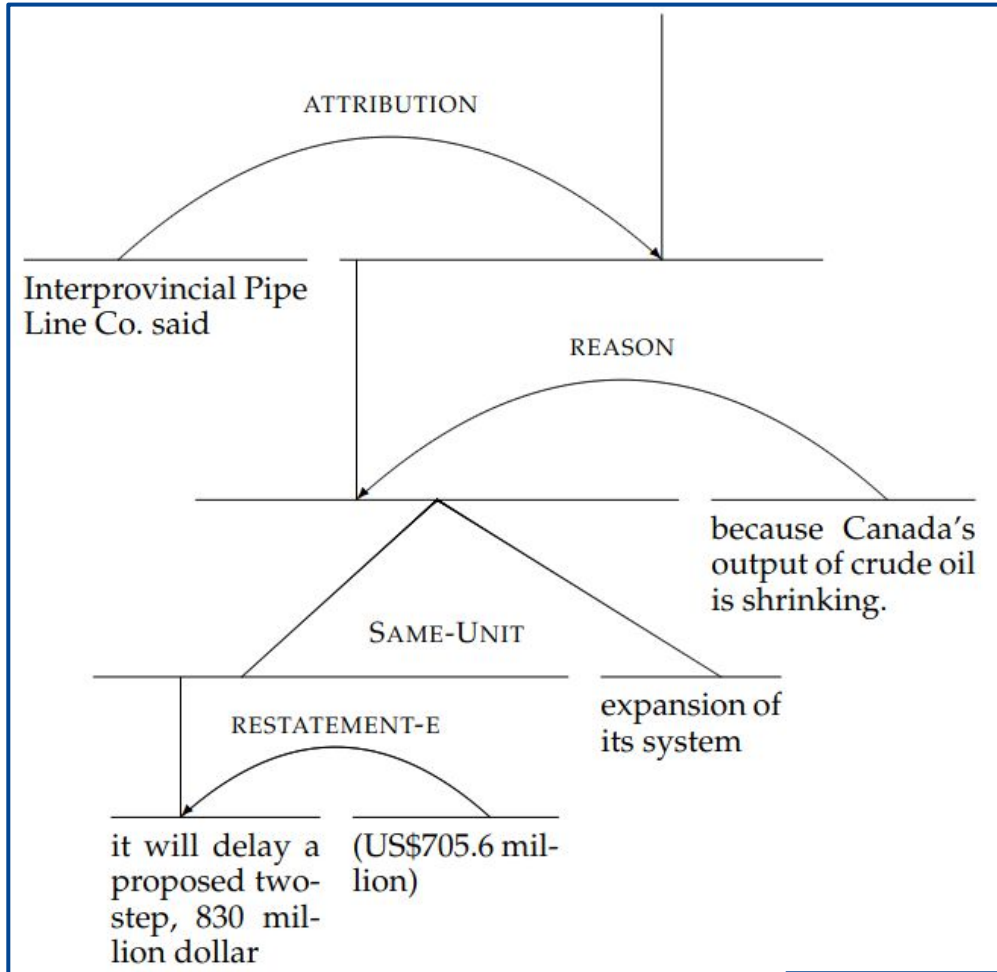
→taken into account in a limited way in main document-level applications (MT, Summarization..)

ANITI Discourse parsing: discourse structure analysis

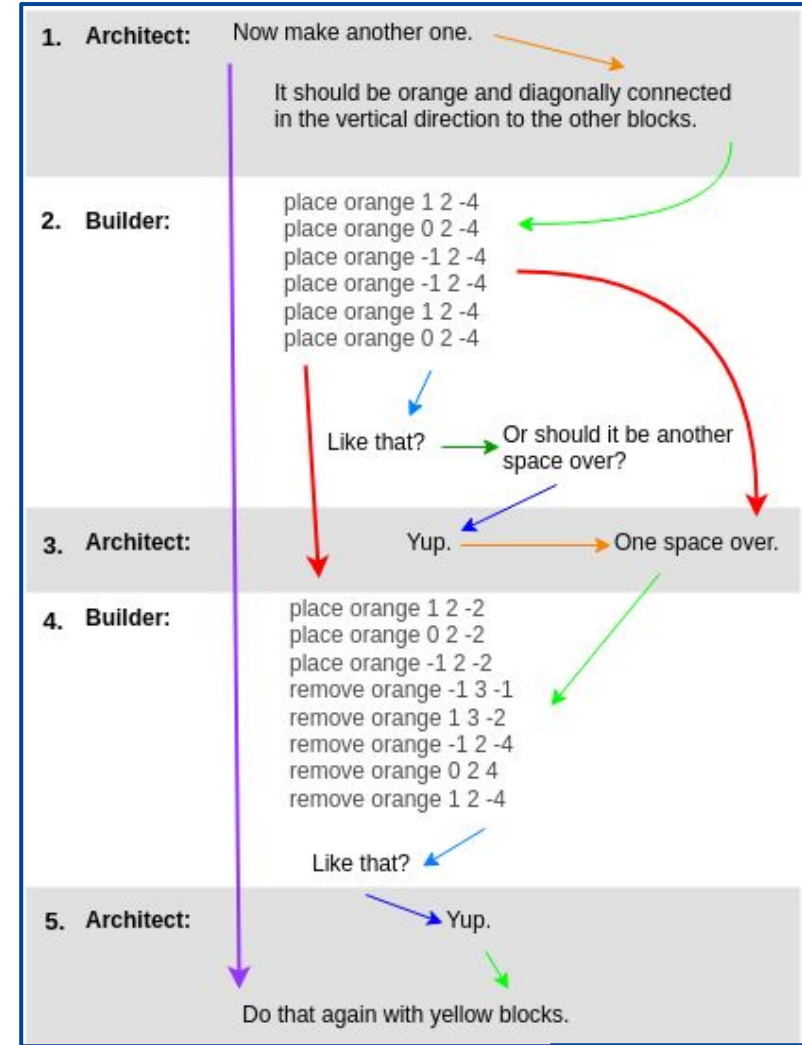
Usually, 3 subtasks:

- (1) **Segmentation**: determining basic relevant units (~clauses)
 - (2) **Attachment/Structure prediction**: determining the overall structure
 - (3) **Labelling**: labelling the structure / prediction relation type
- Task (1): relatively “easy” on written text, harder on conversations
 - Tasks (2)+(3): usually done jointly, much harder

ANITI Discourse parsing: discourse structure analysis



RST DT



Minecraft

ANITI Open problems

- **high-level information**: hard to annotate/provide supervision
- data exist in several **languages**, but with large **disparities**
- **lack of normalization**: competing frameworks, variations across languages, different relation typologies, annotation discrepancies
- involves potentially **large textual contexts**
- in the case of dialogue, generally involve an **extra-linguistic context**, e.g. human-robot collaboration, customer-relationship management ...

ANITI Some approaches from ANITI members

Classic paradigm: fine-tuning of pretrained Language Models (from small to large) → see Kate and Akshay presentation

To overcome some of the mentioned problems :

(1) Transfer:

- (a) between languages e.g. [Metheniti et al. CODI 2024]
- (b) between domains e.g. [Li et al., CODI 2024]
- (c) from written to oral e.g. [Gravellier et al. EMNLP 2021]

(2) Weak supervision:

- (a) exploiting attention matrices in FT PLMs e.g. [Li et al. EACL 2023]
- (b) bootstrapping with weak classifiers (internship)
- (c) data augmentation / generation

ANITI Some approaches from ANITI members

Classic paradigm: fine-tuning of pretrained Language Models (from small to large) → see Kate and Akshay presentation

To overcome some of the mentioned problems :

(1) Transfer:

- (a) between languages e.g. [Metheniti et al. CODI 2024]
- (b) between domains e.g. [Li et al., CODI 2024]
- (c) from written to oral e.g. [Gravellier et al. EMNLP 2021]

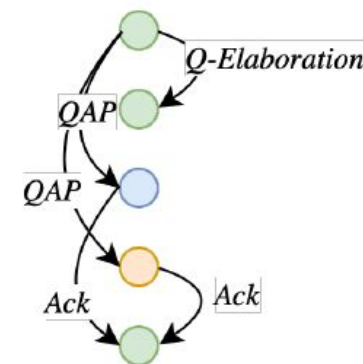
(2) Weak supervision:

- (a) **exploiting attention matrices in FT PLMs** e.g. [Li et al. EACL 2023]
- (b) bootstrapping with weak classifiers (internship)
- (c) data augmentation / generation

Task: predicting document-level structure = directed acyclic graph (text spans and relations)

→ data scarcity for many domains / language

(→ supervised: 20-30% drop cross-domain transfer)

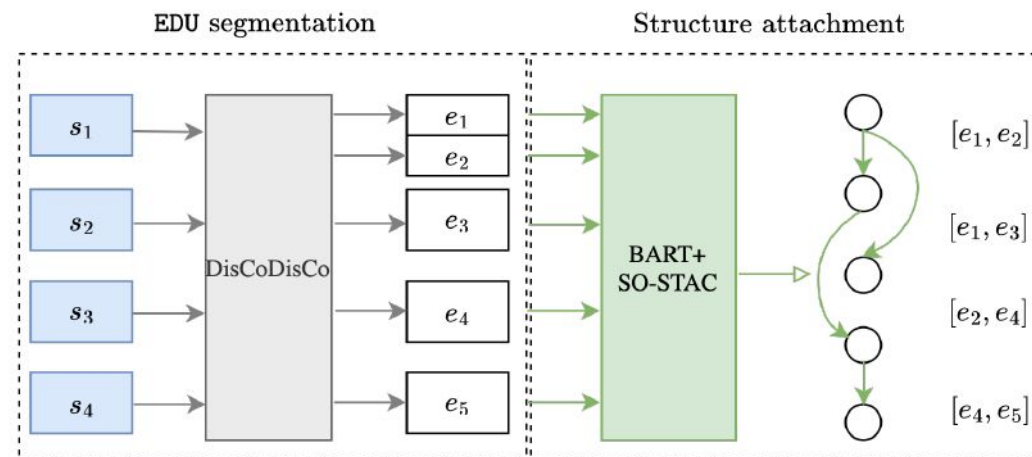


Question: how to extract structure with little supervision?

→ Dialogue structure knowledge in PLMs?

Approach:

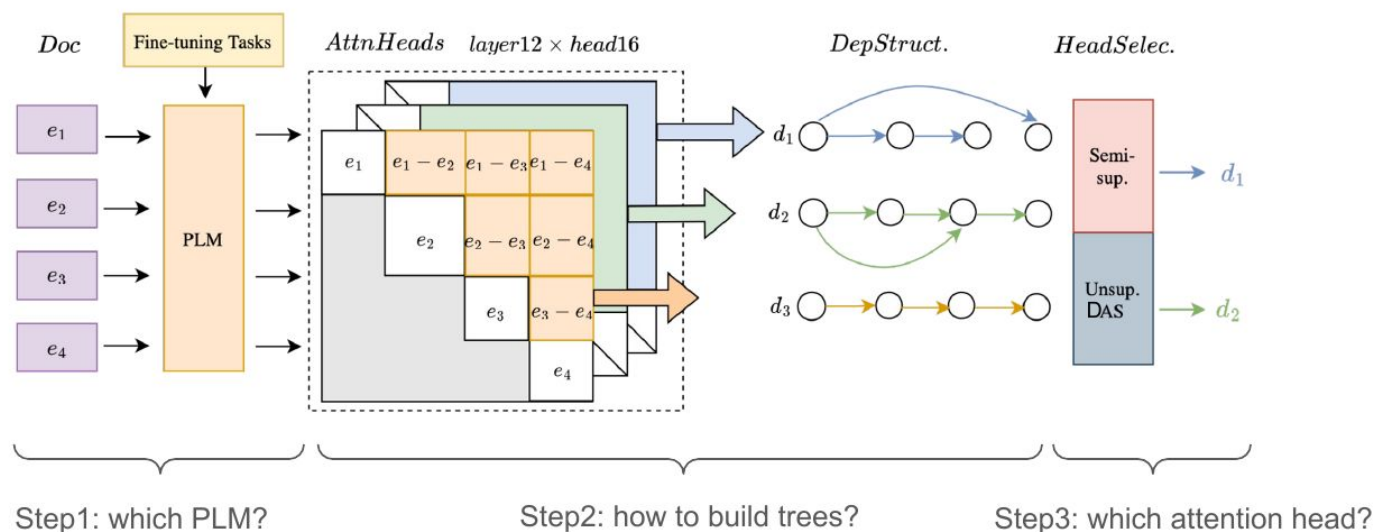
- PLM fine-tuned on related tasks
- use attention scores to extract the structure





Predicting discourse structure with minimal supervision [Li et al. EACL 2023]

- BART FT on sentence ordering in/cross domain, based on speech / speaker info. (best FT task tested)
- Building trees: for each head, use attention score (+Eisner algo.) to compute 'attachment' between text units
- Choose best: unsupervised score or semi-sup. (10-50 examples)



Train on → Test with ↓	BART F ₁	+ SO-DD F ₁	+ SO-STAC F ₁
LAST BSL	56.8	56.8	56.8
Gold H	57.6	58.2	59.5
Unsup H _g	<u>56.6</u>	56.8	56.7
Unsup H _l	56.4	<u>57.1</u>	<u>57.2</u>
Semi-sup 10	57.0 _{0.012}	57.2 _{0.012}	57.1 _{0.026}
Semi-sup 30	57.3 _{0.005}	57.3 _{0.013}	59.2 _{0.009}
Semi-sup 50	57.4_{0.004}	57.7_{0.005}	59.3_{0.007}

ANITI Current work and perspectives

- integrated approaches: multilingual, multi-task, multi-framework
- hierarchical classification to combine frameworks with different relation sets
- weak supervision to combine corpora and models
- extend to other document-level phenomena (e.g. argumentation)
- situated conversational parsing needs better integration with world representation

ANITI References

Metheniti, Eleni, Chloé Braud, and Philippe Muller. "Feature-augmented model for multilingual discourse relation classification." *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*. 2024.

Li, C., Braud, C., Amblard, M., & Carenini, G. (2024, March). Discourse Relation Prediction and Discourse Parsing in Dialogues with Minimal Supervision. In *5th Workshop on Computational Approaches to Discourse* (p. 161).

Li, C., Huber, P., Xiao, W., Amblard, M., Braud, C., & Carenini, G. (2023, May). Discourse Structure Extraction from Pre-Trained and Fine-Tuned Language Models in Dialogues. In *Findings of the Association for Computational Linguistics: EACL 2023*.

Gravellier, Lila, Julie Hunter, Philippe Muller, Thomas Pellegrini, and Isabelle Ferrané. 2021. "Weakly supervised discourse segmentation for multiparty oral conversations." *2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*.

Master internship: Khalil Maachou, Weak supervision for discourse analysis