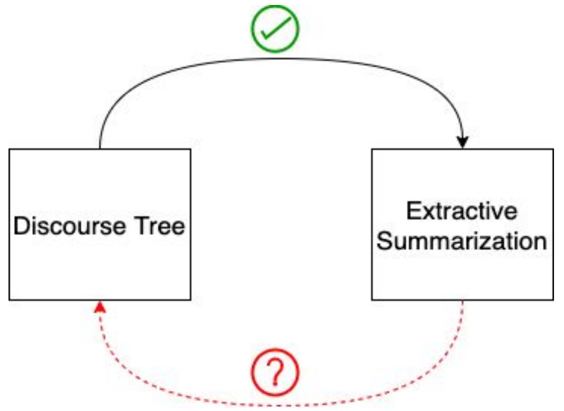
Predicting Discourse Trees from Transformer-based Neural Summarizers

Wen Xiao, Patrick Huber and Giuseppe Carenini
Department of Computer Science, University of British Columbia
xiaowen3@cs.ubc.ca

UBC NLP Group

Motivation

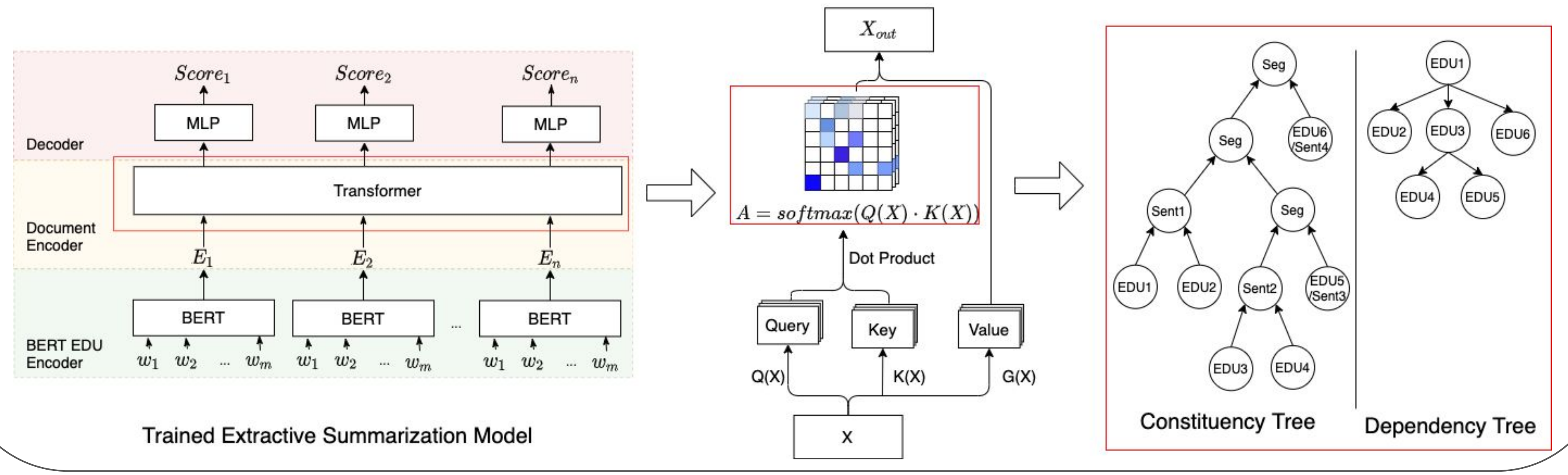
Discourse tree is important for extractive summarization task. [1]



Q: Do Summarizers Learn Discourse Information?

Idea: Does Summarizers' Attention Align with Human-annotated Trees?

1. Build discourse trees based on the attention matrices of trained extractive summarization model,
2. Verify whether and how they are aligned with human-annotated discourse trees.



Conclusion

A: Extractive summarization models do learn discourse information implicitly

- **More dependency information is learnt**
- Most of the discourse information is concentrated **on a single head**.
- The generated trees have **similar properties** as the ground-truth trees, as they both can capture **both local dependencies and long-distance dependencies.**
- The results are consistent across datasets and models → the learned discourse information is **general and transferable inter-domain**.

Step 0: Train Summarizer

Structure:

- BERT EDU Encoder
- Transformer-based Document Encoder
- Decoder

Dataset: CNNDM, NYT

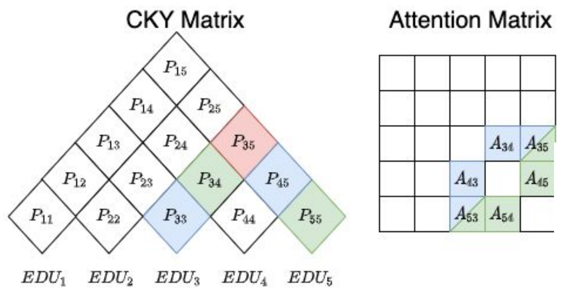
Step 1: Get Attentions

1. Average over each layer
2. Attention matrices per head per layer

Step 2: Build Discourse Trees

Constituency Tree Generation

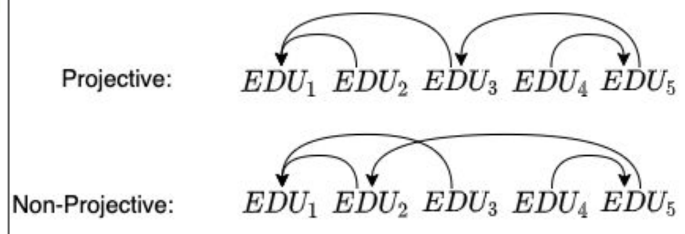
CKY Algorithm: dynamic programming, bottom-up alg.



Dependency Tree Generation

Eisner Algorithm: dynamic programming alg., can only produce projective trees.

CLE Algorithm: find the maximum spanning tree in the graph, and can produce both projective or non-projective trees.



Experiments

Settings

Datasets: RST-DT, Instruction, GUM

Evaluation Metric:

- Constituency Tree: RST-Parseval Score
- Dependency Tree: Unlabeled Attachment Score

Constraints:

- No Constraint / Sentence Constraint

Overall

Model	CKY		Eisner		CLE	
	No Cons.	Sent Cons.	No Cons.	Sent Cons.	No Cons.	Sent Cons.
	RSTDT					
CNNDM-2-1	61.2 / 59.7	76.2 / 74.6	23.7 / 4.8	28.2 / 18.2	21.6 / 1.5	29.3 / 19.6
CNNDM-6-8	60.3 / 60.8	75.4 / 75.0	7.9 / 20.5	13.8 /27.8	7.3 / 17.3	16.1 / 28.5
Random	58.6 (0.1)	74.1 (0.1)	11.2 (0.2)	20.3 (0.2)	1.7 (0.08)	18.7 (0.1)

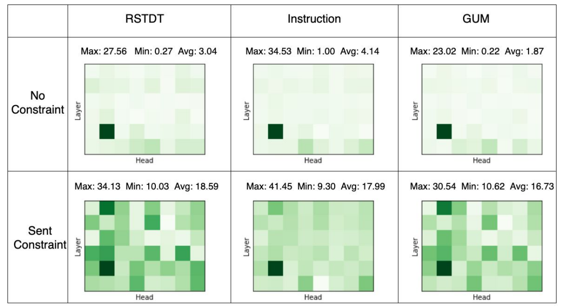
Localness (Best Head)

Measurement(%)	No Cons.	Sent Cons.
RST-DT		
Local Ratio Corr.	77.78	79.17
Instruction		
Local Ratio Corr.	81.15	84.90
GUM		
Local Ratio Corr.	77.99	80.20

Structure Properties (Best Head)

	Branch	Height	Leaf	Arc	vac. (%)
RST-DT					
Ours(No Cons)	1.74	25.76	0.49	0.12	3%
Ground-truth Tree	2.10	8.19	0.51	0.13	2%
Instruction					
Ours(No Cons)	1.80	14.35	0.50	0.14	3%
Ground-truth Tree	1.59	8.49	0.41	0.15	1%
GUM					
Ours(No Cons)	2.14	43.08	0.54	0.08	0%
Ground-truth Tree	2.02	12.17	0.51	0.04	0%

Per-head



* More detailed results and analysis on generated trees can be found in the paper.

[1] Daniel Marcu, Discourse Trees are Good Indicators of Importance in Text, Advances in Automatic Text Summarization (1999)