Predicting Above-Sentence Discourse Structure using Distant Supervision from Topic Segmentation

Patrick Huber*, Linzi Xing* and Giuseppe Carenini

The University of British Columbia

* Equal Contribution

Introduction

"Discourse analysis [...] the analysis of language "beyond the sentence". This contrasts with types of analysis [...] chiefly concerned with the study of grammar"

– Linguistic society of America [Tan12]

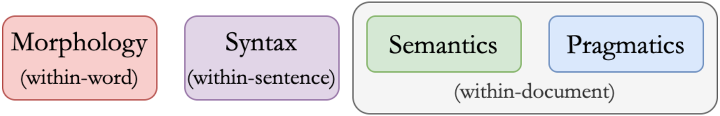
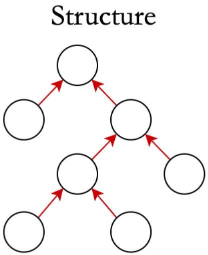


Figure: The spectrum of NLP from small-scale (left) to large-scale (right) structures. Grey box contains mainly discourse-related sub-tasks.

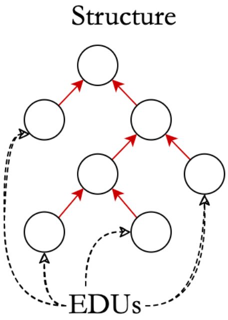
Discourse Analysis

- **Goal:** Reveal structure underlying coherent text
- Structure postulated by discourse theory:
  - **Rhetorical Structure Theory (RST)** [MT88]
  - PDTB [PDL+08]
- RST postulates complete, hierarchical **constituency** trees:



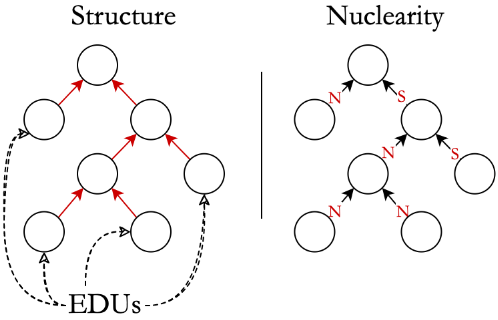
Discourse Analysis

- **Goal:** Reveal structure underlying coherent text
- Structure postulated by discourse theory:
  - **Rhetorical Structure Theory (RST)** [MT88]
  - PDTB [PDL+08]
- RST postulates complete, hierarchical **constituency** trees:



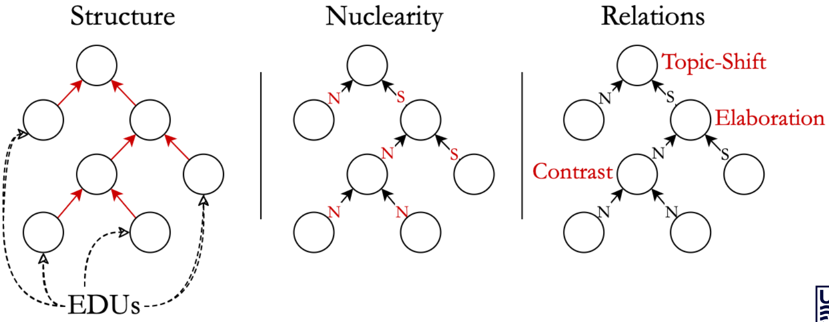
Discourse Analysis

- **Goal:** Reveal structure underlying coherent text
- Structure postulated by discourse theory:
  - **Rhetorical Structure Theory (RST)** [MT88]
  - PDTB [PDL+08]
- RST postulates complete, hierarchical **constituency** trees:



Discourse Analysis

- **Goal:** Reveal structure underlying coherent text
- Structure postulated by discourse theory:
  - **Rhetorical Structure Theory (RST)** [MT88]
  - PDTB [PDL+08]
- RST postulates complete, hierarchical **constituency** trees:

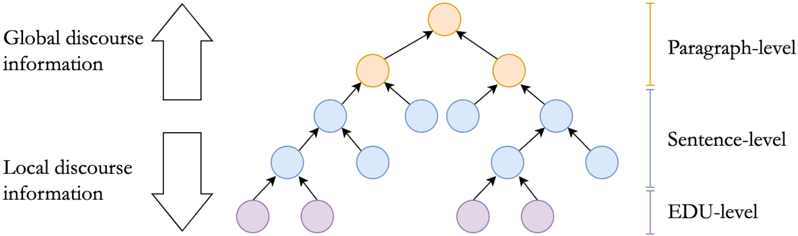


Motivation

Predicting [Above-Sentence Discourse Structure] using [Distant Supervision] from [Topic Segmentation]

Motivation

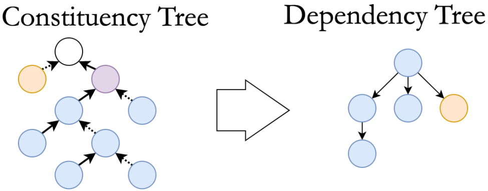
Predicting **[Above-Sentence Discourse Structure]** using [Distant Supervision] from [Topic Segmentation]



* We use the term "paragraph" loosely, including to what is elsewhere called sections

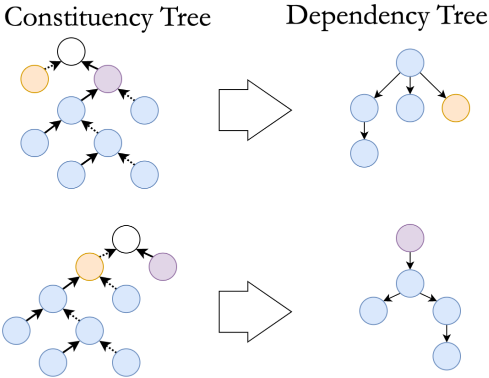
Motivation

Predicting **[Above-Sentence Discourse Structure]** using [Distant Supervision] from [Topic Segmentation]



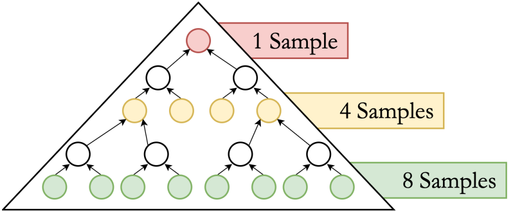
Motivation

Predicting **[Above-Sentence Discourse Structure]** using [Distant Supervision] from [Topic Segmentation]



Motivation

Predicting [Above-Sentence Discourse Structure] using **[Distant Supervision]** from [Topic Segmentation]



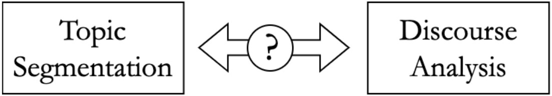
- Variety of downstream tasks shown useful to infer discourse
  - Sentiment analysis → Local structure [HC20]
  - Summarization → Nuclearity [XHC21]
  - What about high-level structures?

Motivation

Predicting [Above-Sentence Discourse Structure] using [Distant Supervision] from **[Topic Segmentation]**

"...long stretches of running text can sensibly be broken into smaller segments [...] motivated by their dealing with a common topic."

– Discourse processing (Book) [Ste11]



Topic Segmentation

Topic segmentation aims to reveal the underlying document structure by splitting documents into topical-coherent textual units.

Example: A Wikipedia article about *City Marcus*

Preface:
Marcus is a city in Cherokee County, Iowa, United States.

History
The first building in Marcus was erected in 1871.
Marcus was incorporated on May 15, 1882.

Geography
Marcus is located at (42.822892, -95.804894).
According to the United States Census Bureau, the city has a total area of 1.54 square miles, all land.

Demographics
As of the census of 2010, there were 1,117 people, 494 households, and 310 families residing in the city.
The population density was 725.3 inhabitants per square mile (280.0/km^2).

......

Topic Segmentation

Topic segmentation aims to reveal the underlying document structure by splitting documents into topical-coherent textual units.

Example: A Wikipedia article about *City Marcus*

Preface:
Marcus is a city in Cherokee County, Iowa, United States.

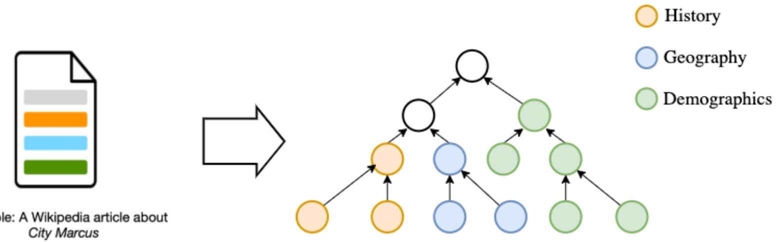
History
The first building in Marcus was erected in 1871.
Marcus was incorporated on May 15, 1882.

Geography
Marcus is located at (42.822892, -95.804894).
According to the United States Census Bureau, the city has a total area of 1.54 square miles, all land.

Demographics
As of the census of 2010, there were 1,117 people, 494 households, and 310 families residing in the city.
The population density was 725.3 inhabitants per square mile (280.0/km^2).

......

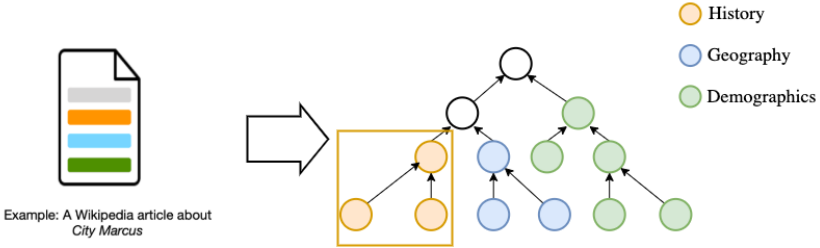
Topic Segmentation



Example: A Wikipedia article about
City Marcus

Assumption: Sentences belong to the same segment are supposed to be more likely merged into a sub-tree on the relatively bottom layer of the discourse tree.

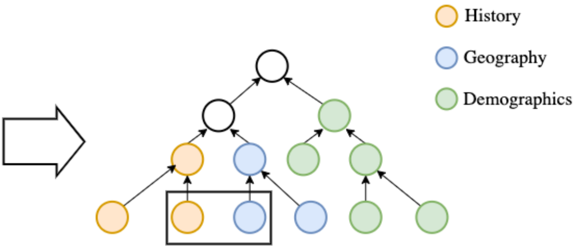
Topic Segmentation



Assumption: Sentences belong to the same segment are supposed to be more likely merged into a sub-tree on the relatively bottom layer of the discourse tree.

Topic Segmentation

Example: A Wikipedia article about *City Marcus*



Assumption: Sentences belong to the same segment are supposed to be more likely merged into a sub-tree on the relatively bottom layer of the discourse tree.

Topic Segmentation

Supervised vs. Unsupervised?

Topic Segmentation

Supervised vs. Unsupervised?

- For monologue text (as the discourse treebank we test on here), large-scale training data is available.

Topic Segmentation

Supervised vs. Unsupervised?

- For monologue text (as the discourse treebank we test on here), large-scale training data is available.
- Better and more robust performance compared to unsupervised methods.

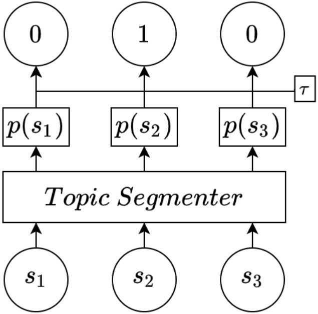
Topic Segmentation

Supervised vs. Unsupervised?

- For monologue text (as the discourse treebank we test on here), large-scale training data is available.
- Better and more robust performance compared to unsupervised methods.

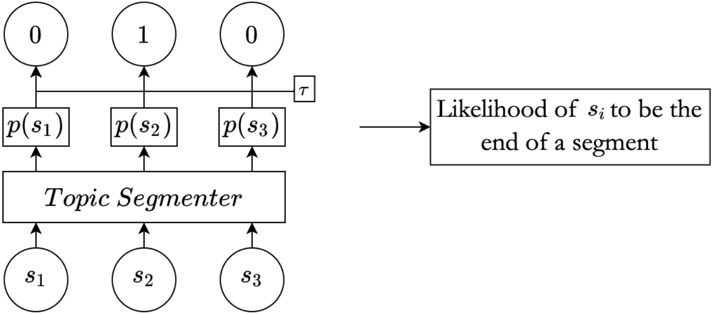
We use the top-performing **supervised** topic segmentation model [XHCT20] to generate discourse structures.

Topic Segmentation



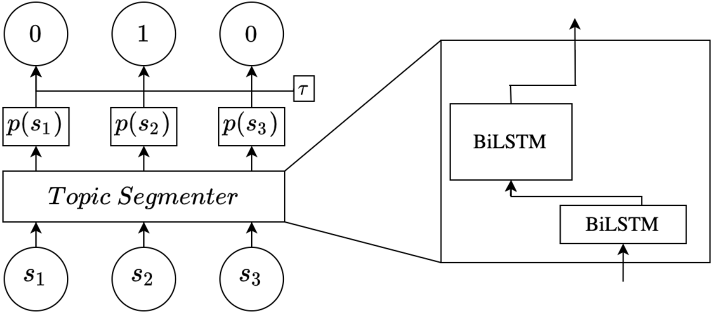
- Binary sequence labelling task

Topic Segmentation



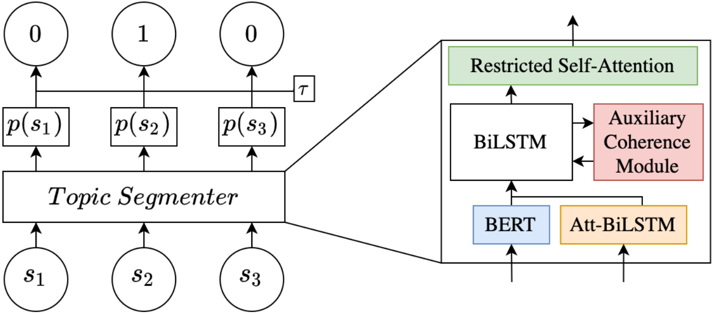
- Binary sequence labelling task

Topic Segmentation



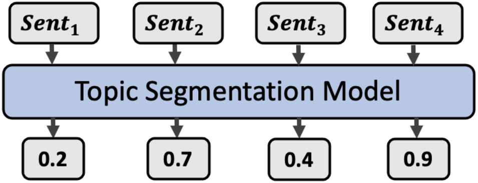
- Binary sequence labelling task
- Basic Model: Hierarchical Bi-LSTM Network [KCM+18]

Topic Segmentation

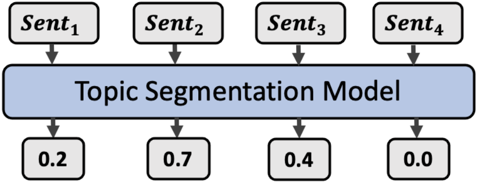


- Binary sequence labelling task
- Basic Model: Hierarchical Bi-LSTM Network [KCM+18]
- Top-performing approach with coherence module [XHCT20]

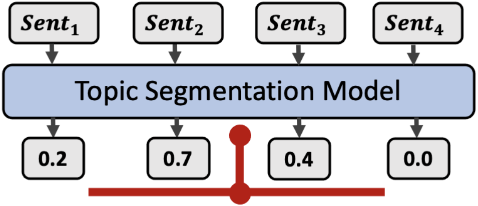
Discourse Tree Generation



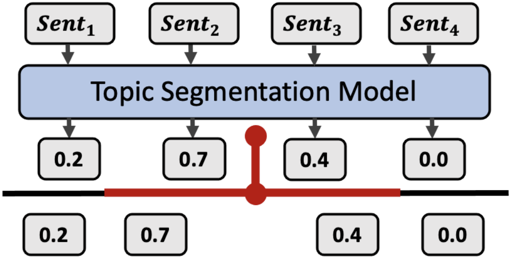
Discourse Tree Generation



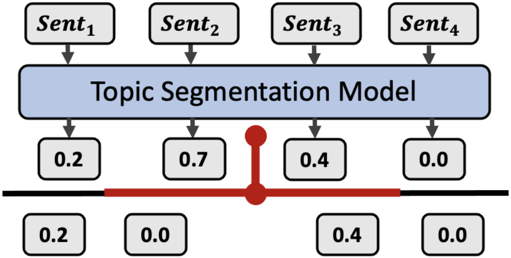
Discourse Tree Generation



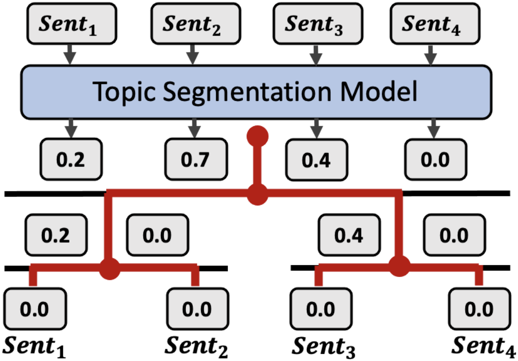
Discourse Tree Generation



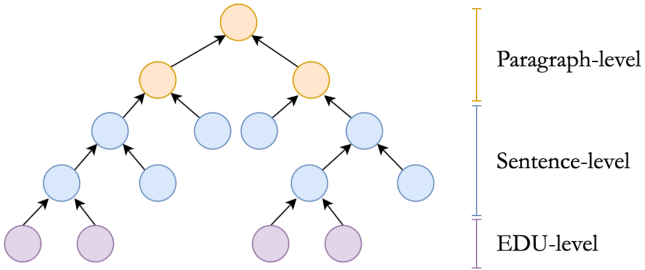
Discourse Tree Generation



Discourse Tree Generation

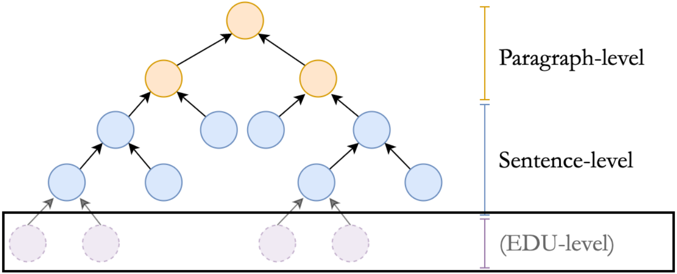


Evaluation Approach



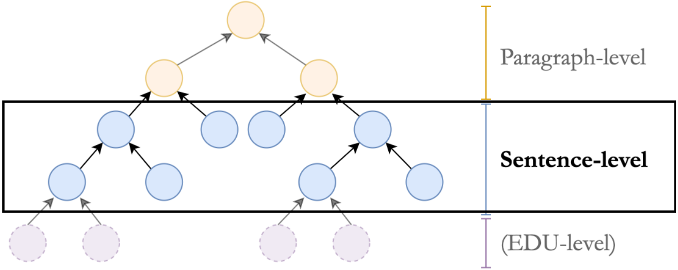
- 3 “natural" document levels

Evaluation Approach



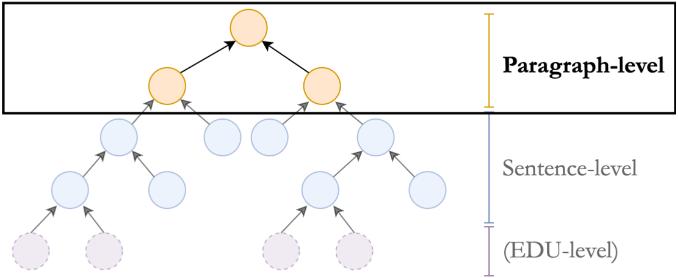
- 3 “natural" document levels
- Topic segmentation operates on sentence-level

Evaluation Approach



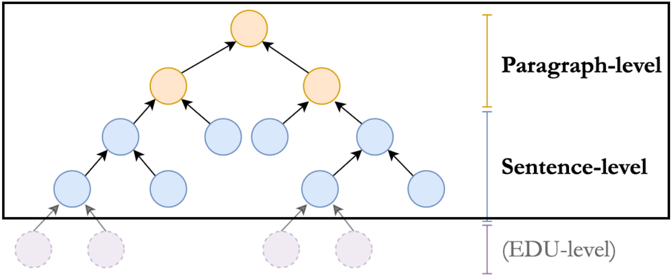
- 3 “natural" document levels
- Topic segmentation operates on sentence-level
- Sentence-level to paragraph-level sub-trees (S-P)

Evaluation Approach



- 3 “natural" document levels
- Topic segmentation operates on sentence-level
- Sentence-level to paragraph-level sub-trees (S-P)
- Paragraph-level to document-level sub-trees (P-D)

Evaluation Approach



- 3 “natural" document levels
- Topic segmentation operates on sentence-level
- Sentence-level to paragraph-level sub-trees (S-P)
- Paragraph-level to document-level sub-trees (P-D)
- Sentence-level to document-level sub-trees (S-D)

Datasets

Dataset	Wikipedia	RST-DT [COM02]	GUM [Zel17]
# of Docs.	20,000	385	150
# of Para./Doc.	31.1	9.99	12.3
# of Sents./Doc.	144.9	22.5	49.3

- Wikipedia Dataset:
  - Randomly sampled from Wikipedia dump
  - Same size as Wiki-Section [ASCM+19], but without domain limitation

Datasets

Dataset	Wikipedia	RST-DT [COM02]	GUM [Zel17]
# of Docs.	20,000	385	150
# of Para./Doc.	31.1	9.99	12.3
# of Sents./Doc.	144.9	22.5	49.3

- Wikipedia Dataset:
  - Randomly sampled from Wikipedia dump
  - Same size as Wiki-Section [ASCM+19], but without domain limitation
- RST-DT Treebank [COM02]:
  - Largest English RST-style discourse treebank (news-domain)
  - Used for training (plain data) and evaluation (tree-structures)

Datasets

Dataset	Wikipedia	RST-DT [COM02]	GUM [Zel17]
# of Docs.	20,000	385	150
# of Para./Doc.	31.1	9.99	12.3
# of Sents./Doc.	144.9	22.5	49.3

- Wikipedia Dataset:
  - Randomly sampled from Wikipedia dump
  - Same size as Wiki-Section [ASCM+19], but without domain limitation
- RST-DT Treebank [COM02]:
  - Largest English RST-style discourse treebank (news-domain)
  - Used for training (plain data) and evaluation (tree-structures)
- GUM Treebank [Zel17]:
  - Multi-domain RST-style discourse treebank
  - Used for training (plain data) and evaluation (tree-structures)

Evaluation Results – Quantitative

Model	RST-DT S-P	RST-DT P-D	RST-DT S-D	GUM S-P	GUM P-D	GUM S-D
Baselines						
Random	77.11	63.90	60.20	67.53	60.96	57.99
Right-Branching	73.57	65.50	59.46	64.15	72.71	59.39
Left-Branching	72.41	64.07	58.07	62.07	54.35	51.56
Supervised RST-style Parsers						
Two-Stage$_{\text{RST-DT}}$	90.64	68.09	72.11	74.20	63.29	63.65
Two-Stage$_{\text{GUM}}$	88.82	65.63	69.58	76.70	**72.94**	**68.38**
SpanBERT$_{\text{RST-DT}}$	**90.75**	**76.03**	**77.19**	–	–	–
Distantly Supervised RST-style Parsers						
Sum$_{\text{CNN/DM}}$	74.23	66.15	59.10	67.89	57.80	53.82
Two-Stage$_{\text{MEGA-DT}}$	85.00	65.50	66.99	73.37	69.88	64.69
TS$_{\text{RST-DT}}$	84.34	62.52	65.96	72.54	67.60	62.79
TS$_{\text{Wiki}}$	83.43	69.78	68.13	**76.98**	63.53	65.84
TS$_{\text{Wiki+RST-DT}}$	83.84	66.54	65.84	–	–	–
TS$_{\text{Wiki+GUM}}$	–	–	–	74.48	67.29	64.69
Ablation – TS$_{\text{Wiki}}$	83.51	68.61	67.47	75.94	64.71	65.38

Table: RST Parseval micro-average precision measure. Best performance per sub-table underlined, best performance per column **bold**.

Evaluation Results – Quantitative

Model	RST-DT S-P	RST-DT P-D	RST-DT S-D	GUM S-P	GUM P-D	GUM S-D
Baselines						
Random	77.11	63.90	60.20	67.53	60.96	57.99
Right-Branching	73.57	65.50	59.46	64.15	72.71	59.39
Left-Branching	72.41	64.07	58.07	62.07	54.35	51.56
Supervised RST-style Parsers						
Two-Stage$_{\text{RST-DT}}$	90.64	68.09	72.11	74.20	63.29	63.65
Two-Stage$_{\text{GUM}}$	88.82	65.63	69.58	76.70	**72.94**	**68.38**
SpanBERT$_{\text{RST-DT}}$	**90.75**	**76.03**	**77.19**	–	–	–
Distantly Supervised RST-style Parsers						
Sum$_{\text{CNN/DM}}$	74.23	66.15	59.10	67.89	57.80	53.82
Two-Stage$_{\text{MEGA-DT}}$	85.00	65.50	66.99	73.37	69.88	64.69
TS$_{\text{RST-DT}}$	84.34	62.52	65.96	72.54	67.60	62.79
TS$_{\text{Wiki}}$	83.43	69.78	68.13	**76.98**	63.53	65.84
TS$_{\text{Wiki+RST-DT}}$	83.84	66.54	65.84	–	–	–
TS$_{\text{Wiki+GUM}}$	–	–	–	74.48	67.29	64.69
Ablation – TS$_{\text{Wiki}}$	83.51	68.61	67.47	75.94	64.71	65.38

Table: RST Parseval micro-average precision measure. Best performance per sub-table underlined, best performance per column **bold**.

Evaluation Results – Quantitative

Model	RST-DT			GUM		
	S-P	P-D	S-D	S-P	P-D	S-D
Baselines						
Random	77.11	63.90	60.20	67.53	60.96	57.99
Right-Branching	73.57	65.50	59.46	64.15	72.71	59.39
Left-Branching	72.41	64.07	58.07	62.07	54.35	51.56
Supervised RST-style Parsers						
Two-Stage$_{\text{RST-DT}}$	90.64	68.09	72.11	74.20	63.29	63.65
Two-Stage$_{\text{GUM}}$	88.82	65.63	69.58	76.70	**72.94**	**68.38**
SpanBERT$_{\text{RST-DT}}$	**90.75**	**76.03**	**77.19**	–	–	–
Distantly Supervised RST-style Parsers						
Sum$_{\text{CNN/DM}}$	74.23	66.15	59.10	67.89	57.80	53.82
Two-Stage$_{\text{MEGA-DT}}$	85.00	65.50	66.99	73.37	69.88	64.69
TS$_{\text{RST-DT}}$	84.34	62.52	65.96	72.54	67.60	62.79
TS$_{\text{Wiki}}$	83.43	69.78	68.13	**76.98**	63.53	65.84
TS$_{\text{Wiki+RST-DT}}$	83.84	66.54	65.84	–	–	–
TS$_{\text{Wiki+GUM}}$	–	–	–	74.48	67.29	64.69
Ablation – TS$_{\text{Wiki}}$	83.51	68.61	67.47	75.94	64.71	65.38

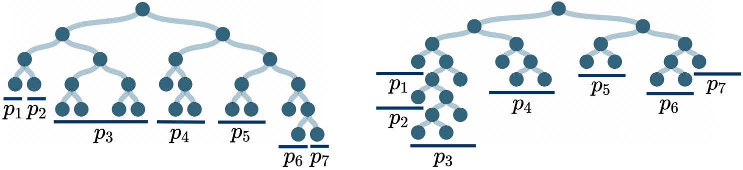
Table: RST Parseval micro-average precision measure. Best performance per sub-table underlined, best performance per column **bold**.

Evaluation Results – Quantitative

Genre	Right-Branching	Two-Stage (GUM)	TS (Wiki)
Travel guides	**78.1**	75.0	53.1
Biographies	75.0	**78.6**	**78.6**
Fiction	**80.6**	**80.6**	61.1
How-to guides	**69.4**	64.3	66.3
Academic writing	70.4	**81.5**	70.4
News stories	57.4	57.4	**63.2**
Political speeches	80.0	**85.0**	60.0
Textbooks	**78.6**	71.4	57.1
Interviews	78.8	**83.3**	60.6

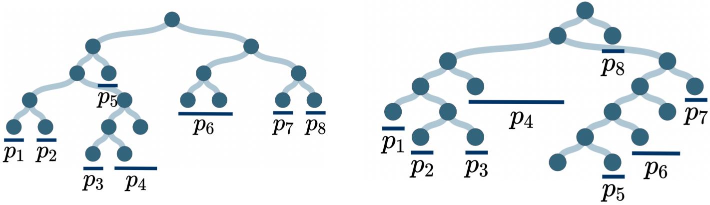
- Similar domains to Wikipedia reach the best performance
- Right-branching structures strong baseline

Evaluation Results – Qualitative



- Prediction (left) according to topic segment probabilities
- Gold-standard (right) from RST-DT corpus
- Showcase open problem:
  - “Nested paragraphs”

Evaluation Results – Qualitative



- Prediction (left) according to topic segment probabilities
- Gold-standard (right) from RST-DT corpus
- Showcase open problem:
  - "Nested paragraphs"
  - "Long-distance paragraphs"

Conclusion and Future Work

- Topic segmentation provides useful signals for high-level discourse constituency trees
- Greedy top-down algorithm performs well on RST-DT and GUM
- Giving insights into tree structure prediction based on textual levels

- Investigate non-greedy tree aggregation, e.g., CKY
- Incorporate discourse signals into topic segmentation models
- Use dense representations of neural topic segmenters to infer discourse structures with nuclearity and relation labels

References I

- Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser, *Sector: A neural model for coherent topic segmentation and classification*, Transactions of the Association for Computational Linguistics **7** (2019), 169–184.
- Lynn Carlson, Mary Ellen Okurowski, and Daniel Marcu, *RST discourse treebank*, Linguistic Data Consortium, University of Pennsylvania, 2002.
- Patrick Huber and Giuseppe Carenini, *Mega rst discourse treebanks with structure and nuclearity from scalable distant sentiment supervision*, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7442–7457.

References II

- Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant, *Text segmentation as a supervised learning task*, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (New Orleans, Louisiana), Association for Computational Linguistics, June 2018, pp. 469–473.
- William C Mann and Sandra A Thompson, *Rhetorical structure theory: Toward a functional theory of text organization*, Text-Interdisciplinary Journal for the Study of Discourse **8** (1988), no. 3, 243–281.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber, *The penn discourse treebank 2.0.*, LREC (2008).
- Manfred Stede, *Discourse processing*, Synthesis Lectures on Human Language Technologies **4** (2011), no. 3, 1–165.

References III

Deborah Tannen, *Discourse analysis–what speakers do in conversation*, Linguistic society of America (2012).

Wen Xiao, Patrick Huber, and Giuseppe Carenini, *Predicting discourse trees from transformer-based neural summarizers*, arXiv preprint arXiv:2104.07058 (2021).

Linzi Xing, Brad Hackinen, Giuseppe Carenini, and Francesco Trebbi, *Improving context modeling in neural topic segmentation*, Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020, pp. 626–636.

Amir Zeldes, *The GUM corpus: Creating multilayer resources in the classroom*, Language Resources and Evaluation **51** (2017), no. 3, 581–612.