

Unified Sequential Discourse Parsing

Zhang Ying

Written by
Hidetaka Kamigaito and Zhang Ying

EDU Segmentation

Introduction

- A task for splitting a sentence into element discourse units (EDUs)
- The EDU is used for discourse parsing, text summarization, etc.

An input sentence

The Treasury also said noncompetitive tenders will be considered timely if postmarked no later than Sunday, Oct.29, and received no later than tomorrow.



Segmentation

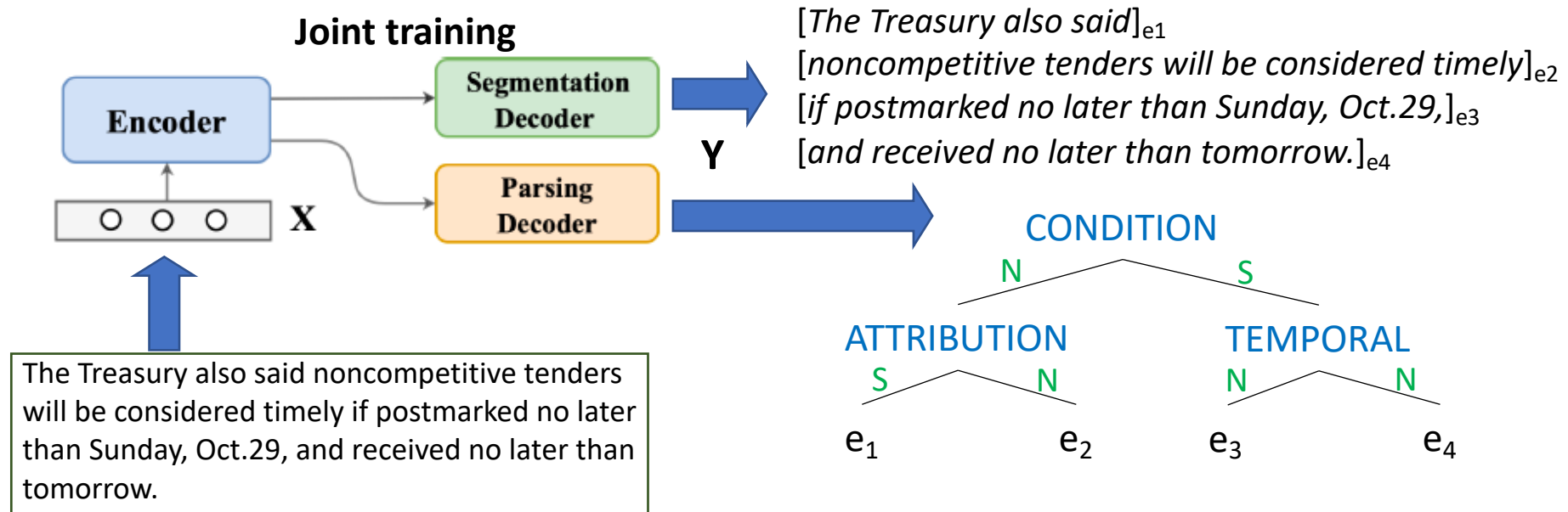
[The Treasury also said]_{e1} [noncompetitive tenders will be considered timely]_{e2} [if postmarked no later than Sunday, Oct.29,]_{e3} [and received no later than tomorrow.]_{e4}

Element Discourse Units (EDUs)

Unified Sentence-Level Discourse Parsing

SoTA segmenter

- Lin+ 2019 achieved the SoTA F_1 score 95.55 by using the same encoder for both parser and segmenter to jointly train them



Unified Sentence-Level Discourse Parsing

Room for improvement

- This approach has two weak points:
 - It does not consider interaction between EDU segmentation and discourse parsing in the decoding step
 - It only considers $P(Y|X)$, conditional probability from an input sentence to the output labels

How to deal with these problems?

Related Work: Towards String-to-Tree NMT

<https://www.aclweb.org/anthology/P17-2021.pdf>

- Aharoni+ 2017 insert constituency tags into target sentences of training data to translate with utilizing **syntactic tree information**

Jane hatte eine Katze . \rightarrow (*ROOT* (*S* (*NP* **Jane**) *NP* (*VP* **had** (*NP* **a cat**) *NP*) *VP* .) *S*) *ROOT*

- Although the increase of a sentence length, BLEU scores are improved in the several test sets.

system	newstest2015	newstest2016
bpe2bpe	27.33	31.19
bpe2tree	27.36	32.13
bpe2bpe ens.	28.62	32.38
bpe2tree ens.	28.7	33.24

Table 1: BLEU results for the WMT16 experiment

	system	newstest2015	newstest2016
DE-EN	bpe2bpe	13.81	14.16
	bpe2tree	14.55	16.13
	bpe2bpe ens.	14.42	15.07
	bpe2tree ens.	15.69	17.21
RU-EN	bpe2bpe	12.58	11.37
	bpe2tree	12.92	11.94
	bpe2bpe ens.	13.36	11.91
	bpe2tree ens.	13.66	12.89
CS-EN	bpe2bpe	10.85	11.23
	bpe2tree	11.54	11.65
	bpe2bpe ens.	11.46	11.77
	bpe2tree ens.	12.43	12.68

Table 2: BLEU results for the low-resource experiments (News Commentary v8)

- By using the same approach, we can consider discourse tree information for discourse segmentation

Related Work: Neural Generative Rhetorical Structure Parsing

<https://www.aclweb.org/anthology/D19-1233.pdf>

- Mabona+ 2019 adopt RNNG to RST parsing
 - Different from the original one, their GEN action generates an EDU instead of a word
- They parsed texts on gold EDUs
 - Didn't try EDU segmentation

Action	Before	After	Probability	Condition
GEN(e)	$\langle S, B \rangle$	$\langle S \text{EDU}(e), B e \rangle$	$p_{\text{trans}}(\text{GEN} S) \cdot p_{\text{gen}}(e S)$	$ B < m$
RE(r, n)	$\langle S U_L U_R, B \rangle$	$\langle S (\text{Unit}(r, n) U_L U_R), B \rangle$	$p_{\text{trans}}(\text{RE}(r, n) S)$	$ S \geq 2$

Table 1: Our transition system. $|S|$ is the number of discourse units on the stack, $|B|$ is the number of EDUs in the buffer and m is the number of EDUs in the whole document, r is a relation label and n is a nuclearity label.

Stack	Buffer	Prediction
ϵ	ϵ	GEN(e_1)
EDU(e_1)	e_1	GEN(e_2)
EDU(e_1) EDU(e_2)	$e_1 e_2$	GEN(e_3)
EDU(e_1) EDU(e_2) EDU(e_3)	$e_1 e_2 e_3$	RE(ATTR, SN)
EDU(e_1) (Unit(ATTR, SN) EDU(e_2) EDU(e_3))	$e_1 e_2 e_3$	RE(JUST, NS)
(Unit(JUST, NS) EDU(e_1) (Unit(ATTR, SN) EDU(e_2) EDU(e_3)))	$e_1 e_2 e_3$	

[e_1 Acme Inc. has closed several widget factories.] [e_2 The CEO told investors] [e_3 they were no longer profitable.]

Table 2: An example of a completed computation in our transition system.

Related Work: Parsing as Language Modeling

<https://www.aclweb.org/anthology/D16-1257.pdf>

- Choe+ 2016 also propose a generative constituency parser based on a reranking approach (importance sampling)
 - They rerank the k-best results of the Charniak parser by utilizing an LSTM language model
- We can use the same idea in EDU segmentation

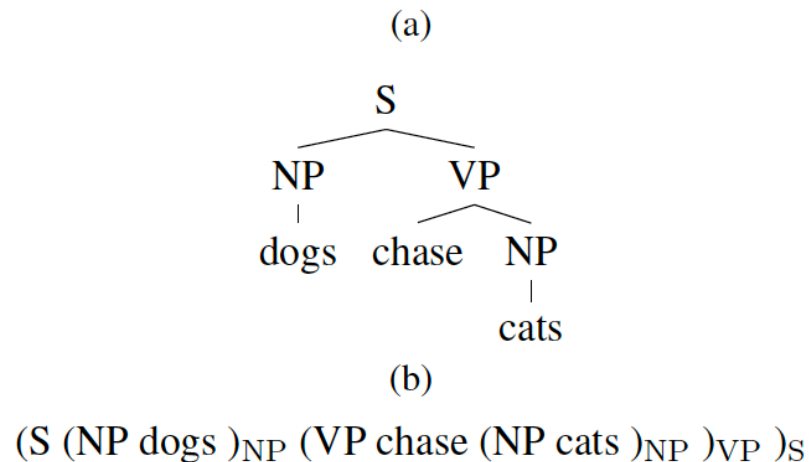
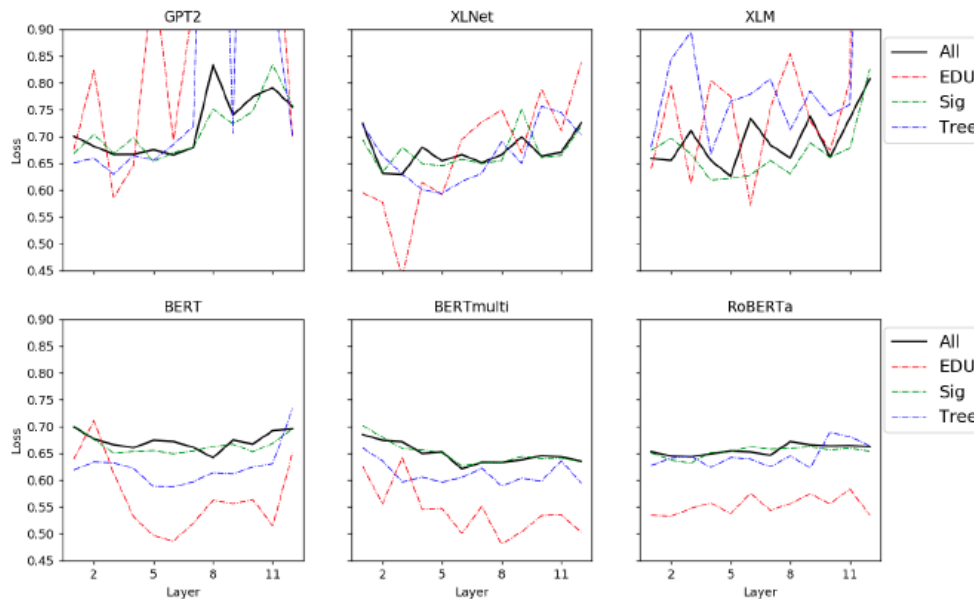


Figure 1: A tree (a) and its sequential form (b). There is a one-to-one mapping between a tree and its sequential form. (Part-of-speech tags are not used.)

Related Work: Rhetorical Capacities of Neural Language Models

<https://arxiv.org/pdf/2010.00153.pdf>

- Zhu+ 2020 adopt language models to RST to investigate the obtained information in the pretrained models
 - Based on frozen features (didn't fine-tune)
 - Their purpose is not to parsing and segmentation



It seems the last layer of multilingual-BERT works well.

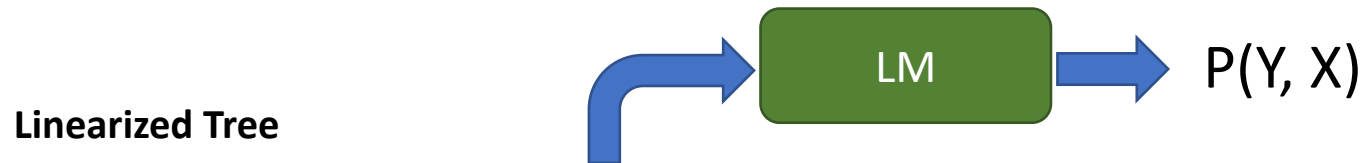
However, the fine-tuning case is still uncertain.

Figure 3: Loss vs layer plot of six neural LMs on four RST feature sets on IMDB. The solid lines represent all RST features combined, while each dash-dotted line denotes one component (EDU, Sig, or Tree feature group for red, green, and blue respectively). In general, BERT-based LMs (BERT, BERT-multi, RoBERTa) encode rhetorical features in a more stable and easy-to-probe manner than the rest.

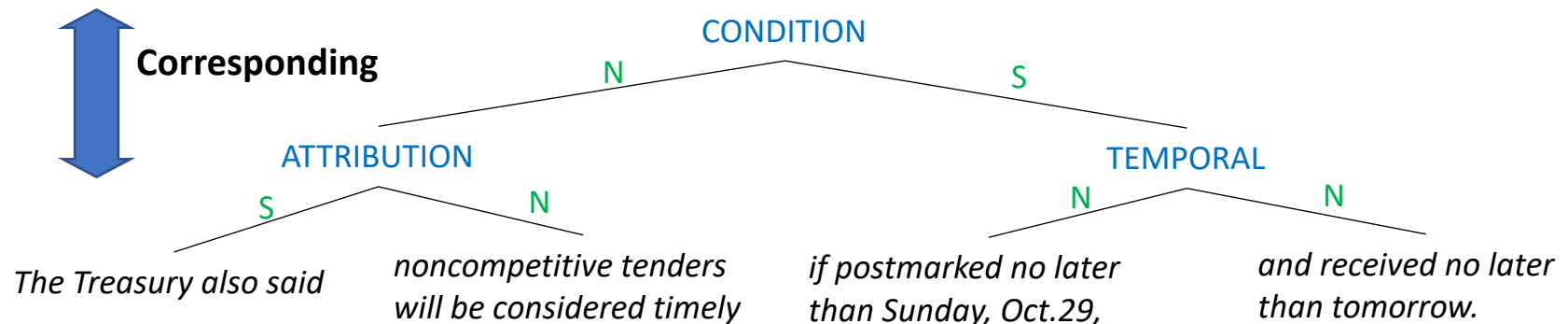
Proposal: Unified Sequential Discourse Parsing

Overview (1)

- Predicting joint probability $P(Y, X)$ for a linearized discourse tree of a sentence based on language model
 - Probability is not restricted to the conditional $P(Y|X)$
 - This method can consider interaction between labels of discourse tree and segmentation



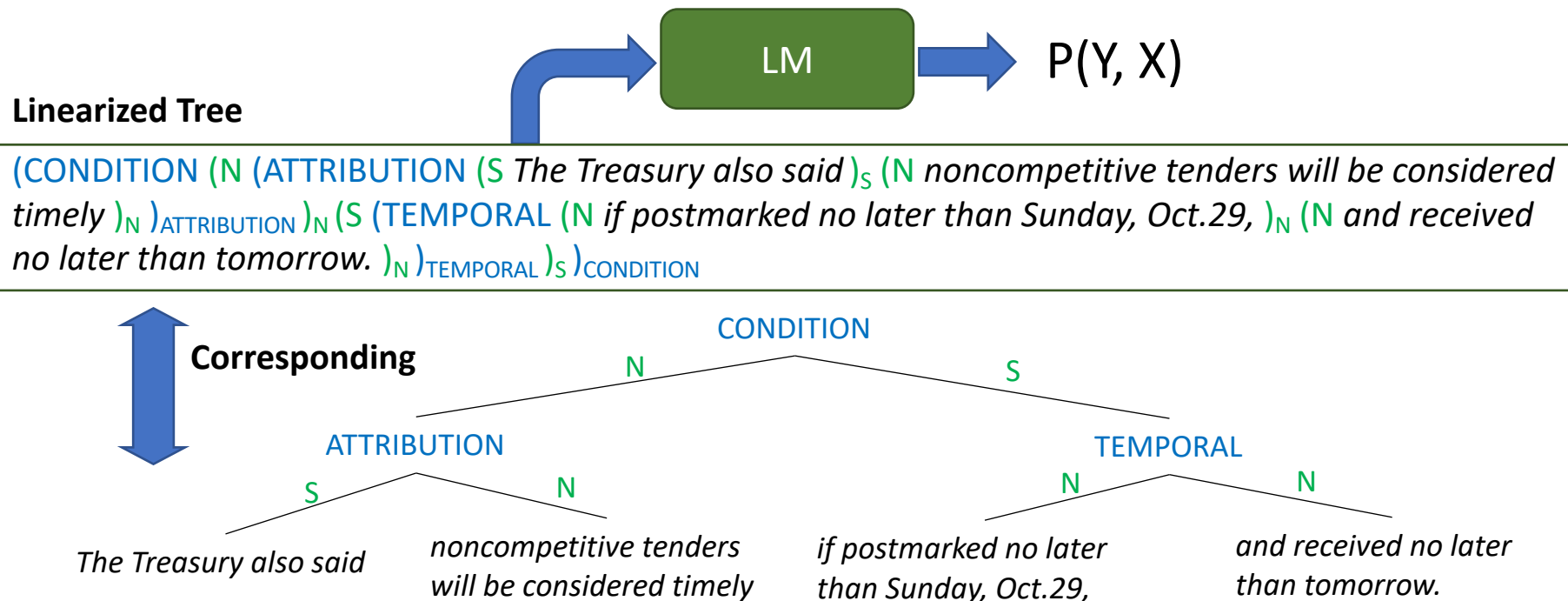
(CONDITION (N (ATtribution (S The Treasury also said)_S (N noncompetitive tenders will be considered timely)_N)_{ATtribution})_N (S (TEMPORAL (N if postmarked no later than Sunday, Oct.29,)_N (N and received no later than tomorrow.)_N)_{TEMPORAL})_S)_{CONDITION}



Proposal: Unified Sequential Discourse Parsing

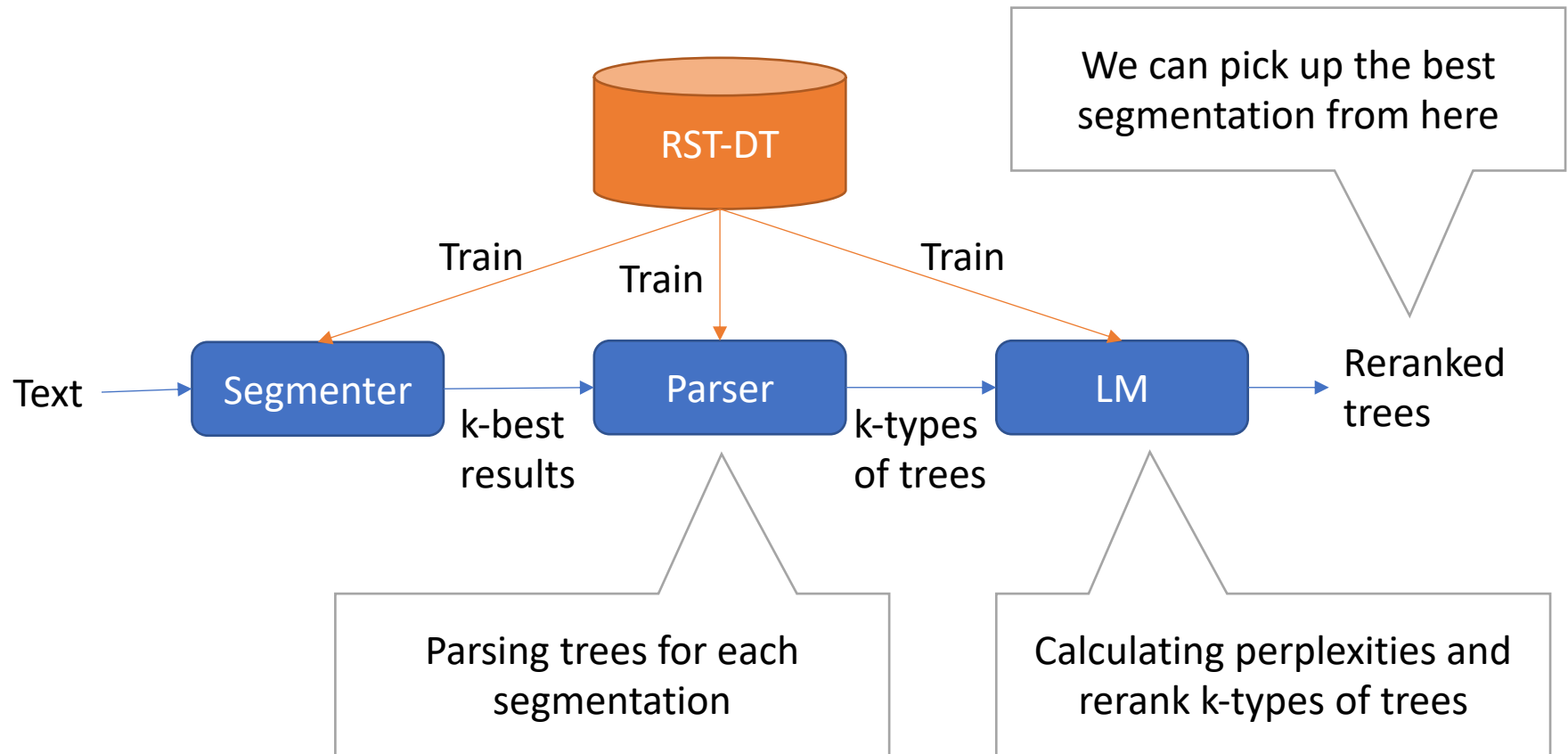
Overview (2)

- However, it **cannot segment texts only based on $P(Y, X)$**
- To deal with this problem, we first generate candidates by the conventional segmenter and then rerank them based on $P(Y, X)$



Segmentation process for the proposed method

Data flow



Additional information for LM

For enhancing label embeddings

- Incorporates definitions for each label name into LM by using **next sentence prediction**
 - Just concatenating the label and its definition

Label	Definition
Condition	the state that something or someone is in
Attribution	the act of saying or thinking that something is the result or work of a particular person or thing
Temporal	relating to practical matters or physical things, rather than spiritual ones

⋮

Comparison methods

Baseline and proposed methods

- Baseline: The segmenter used for generating candidates
- Proposed methods:

- With all labels

(CONDITION (N (ATtribution (S The Treasury also said)S (N noncompetitive tenders will be considered timely)N)ATtribution)N (S (TEMPORAL (N if postmarked no later than Sunday, Oct.29,)N (N and received no later than tomorrow.)N)TEMPORAL)S)CONDITION

- With nuclearity labels

(N (S The Treasury also said)S (N noncompetitive tenders will be considered timely)N)N (S (N if postmarked no later than Sunday, Oct.29,)N (N and received no later than tomorrow.)N)S

- With relation labels

(CONDITION (ATtribution The Treasury also said | noncompetitive tenders will be considered timely)ATtribution (TEMPORAL if postmarked no later than Sunday, Oct.29, | and received no later than tomorrow.)TEMPORAL)CONDITION

- Without labels

The Treasury also said | noncompetitive tenders will be considered timely | if postmarked no later than Sunday, Oct.29, | and received no later than tomorrow.

Evaluation Metric

Segmentation and Parsing

- We can expect improvement for both segmentation and parsing performance
 - Thus, we should use the metric for both discourse parsing and segmentation
- Segmentation: Micro-average- F_1 for labels
- Parsing: Micro-average- F_1 for Span, Nuclearity, Relation, and Full
- Significance: Paired Bootstrap Resampling. Repeatedly create 1000 new virtual test sets. If one system outperforms the other system 95% of the time, we draw the conclusion that it is better with 95% statistical significance. (Koehn, 2004)

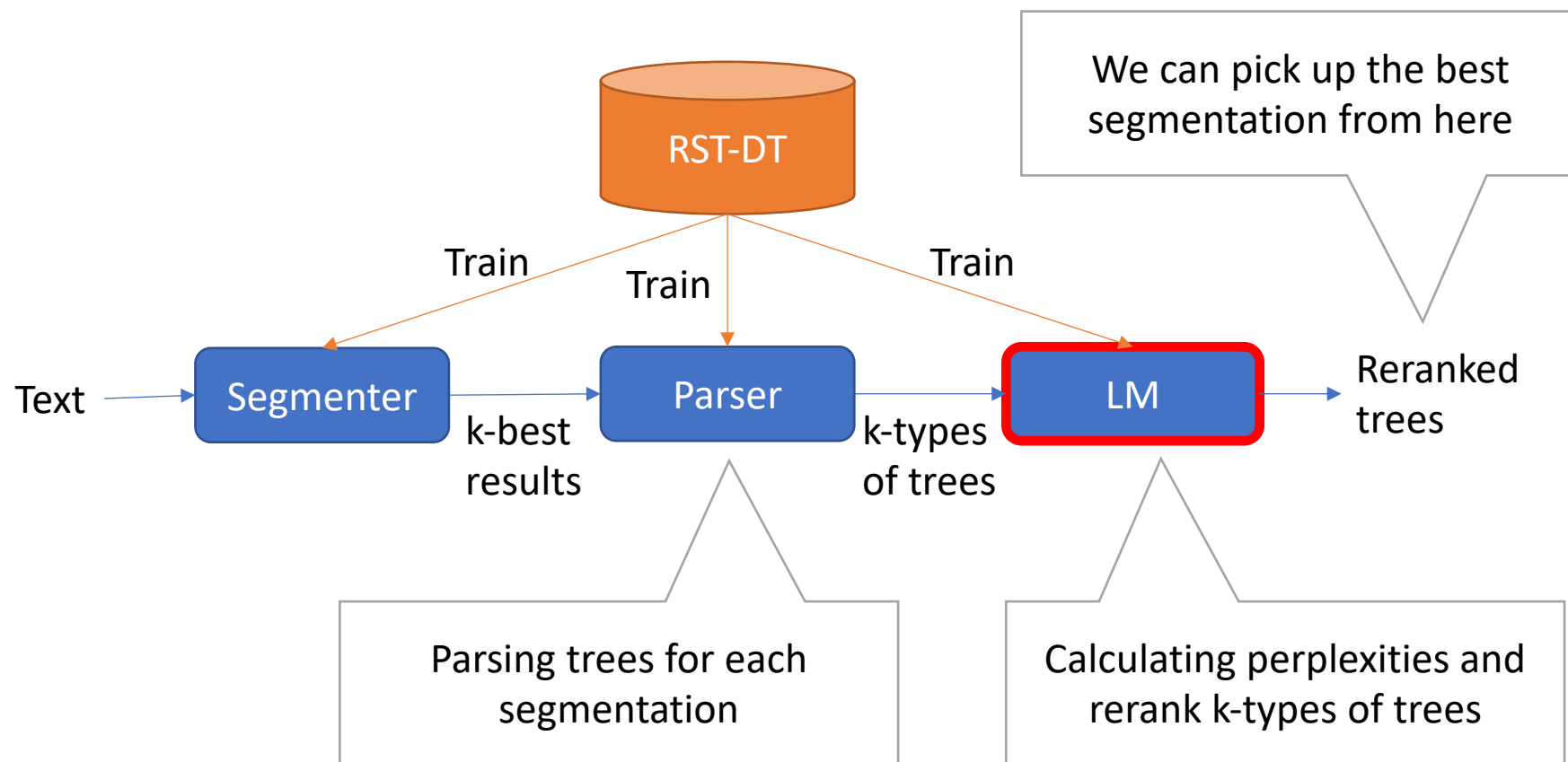
Current Progress

- Segmenter
 - Run a baseline EDU segmenter
 - Rerank the results with a language model
 - Evaluate the result
- Sentence discourse parser
 - Run several parsers as baselines
 - Rerank the results with a language model
 - Evaluate the result
- Sentence discourse parse on automatic segmentation data
 - Run several parsers as baselines
 - Rerank the results with a language model
 - Evaluate the result

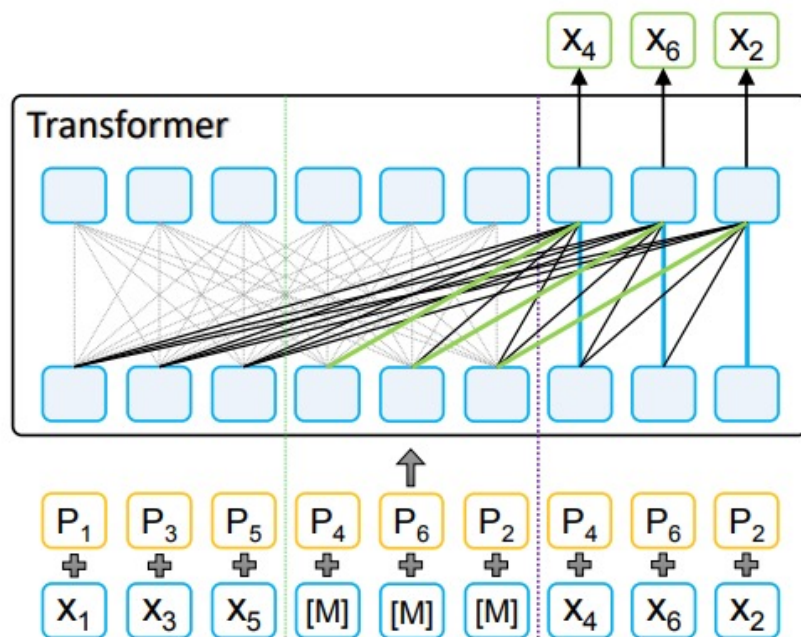
Language Model

Reranker

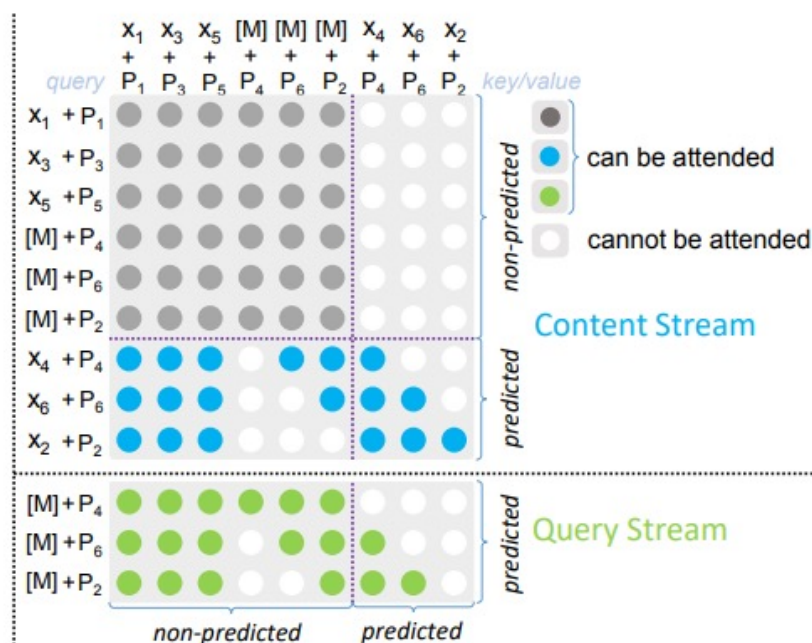
- **MPNet** is used for the language model to rerank the parse results



- MPNet is an extension of BERT and XLNet
 - masked language model + permuted language model



(a)



(b)

- MPNet uses the value of X_4 to predict X_6 , instead of using [MASK] in MLM.
- MPNet uses the position of X_6 to predict X_4 , instead of using no position in PLM.

MPNet: Masked and Permuted Pre-training for Language Understanding

Example of using MPNet to compute the probability of a sentence

- Original sentence

The Treasury also said [EDU] noncompetitive tenders will be considered timely [EDU] if postmarked no later than Sunday, Oct.29, [EDU] and received no later than tomorrow. [EDU]

- During training step

Mask a part (15%) of the input sentence: *The Treasury also said [MASK5] noncompetitive [MASK7] ...*

Get a set of possible discourse parsing: $Y = \{y_{gold}, y_{cand_1}, y_{cand_2}, y_{cand_3}, \dots\}$

$$p(\mathbf{x}, y_{gold}) = p(\mathbf{z}_{gold}) = p(z_1, \dots, z_m) = \prod_{t=1}^m p(z_t | z_1, z_2, \dots, z_{t-1}, z_{t+1}, \dots, z_m)$$

$$p(\mathbf{x}, y_{cand_i}) = p(\mathbf{z}_{cand_i}) = p(z_1, \dots, z_m) = \prod_{t=1}^m (1 - p(z_t | z_1, z_2, \dots, z_{t-1}, z_{t+1}, \dots, z_m))$$


- During test step, mask only one word each time, and use a loop to predict the whole sentence.

Experiment 1: Discourse segmentation without labels

Example of input sentence

- 1. For a given sentence, generate the candidate segmentations by the baseline model.

Elementary Discourse Units (EDU)



The Treasury also said | noncompetitive tenders will be considered timely | if postmarked no later than Sunday, Oct.29, | and received no later than tomorrow. |

- 2. Build input sentence

The Treasury also said [EDU] noncompetitive tenders will be considered timely [EDU] if postmarked no later than Sunday, Oct.29, [EDU] and received no later than tomorrow. [EDU]

- Because MPNet uses the vocab of Bert, token [EDU] is replaced by token [SEP] in our code.

label embeddings with its definition

Example of input sentence with label definition embedding

3. Build input sentence with label embedding. Concatenate the definition with input sentence, or replace the word embedding of input sentence by averaged definition embedding.

<BOS> *The Treasury also said* **[SEP]** *noncompetitive tenders will be considered timely*
[SEP] *if postmarked no later than Sunday, Oct.29,* **[SEP]** *and received no later than*
tomorrow. **[SEP]** **<EOS>**

Definition
(won't be masked)

[SEP] : elementary discourse units are the minimal building blocks of a discourse tree

Experiment 1: Discourse segmentation without labels

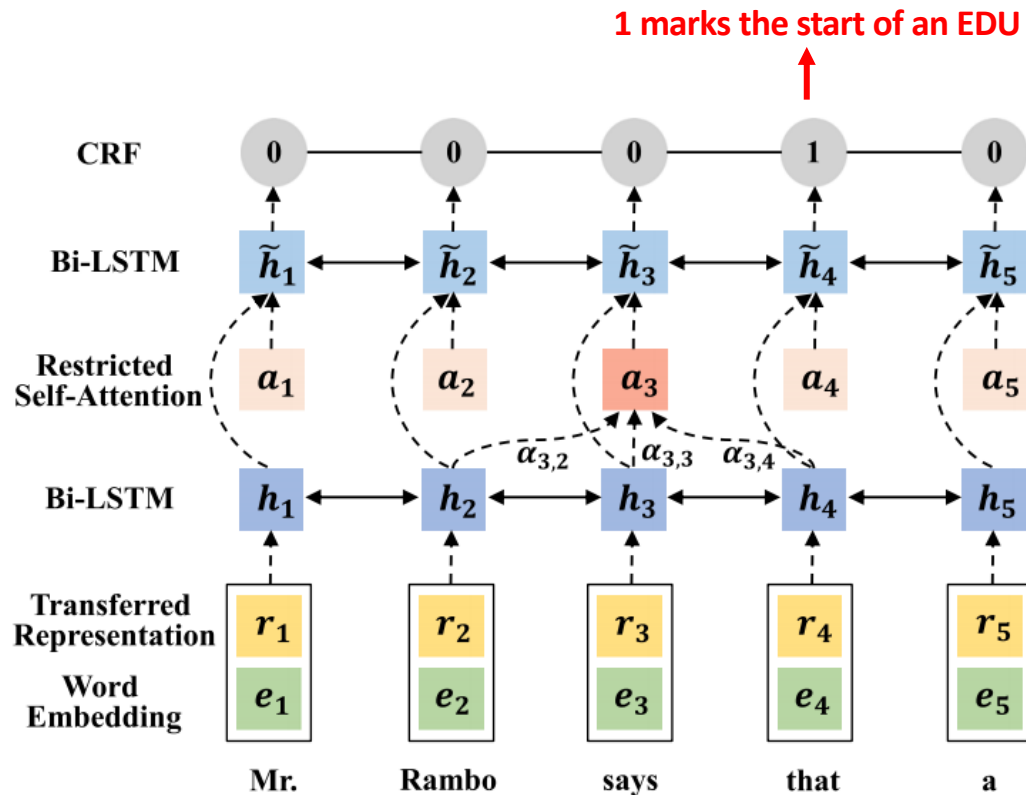
SoTA segmenter: A Unified Linear-Time Framework for Sentence-Level Discourse Parsing

- The author didn't provide codes for data preprocessing.
- In their [github](#), they mentioned the data format of the input sentence should be like followings which contains the segment information, we think this is not suitable for a discriminative model as our baseline.
 - Although the [report,] which has [released] before the stock market [opened,] didn't trigger the 190.58 point drop in the Dow Jones Industrial [Average,] analysts [said] it did play a role in the market's [decline.]
 - '[' denotes the EDU boundary tokens.

Experiment 1: Discourse segmentation without labels

Baseline model: Toward Fast and Accurate Neural Discourse Segmentation

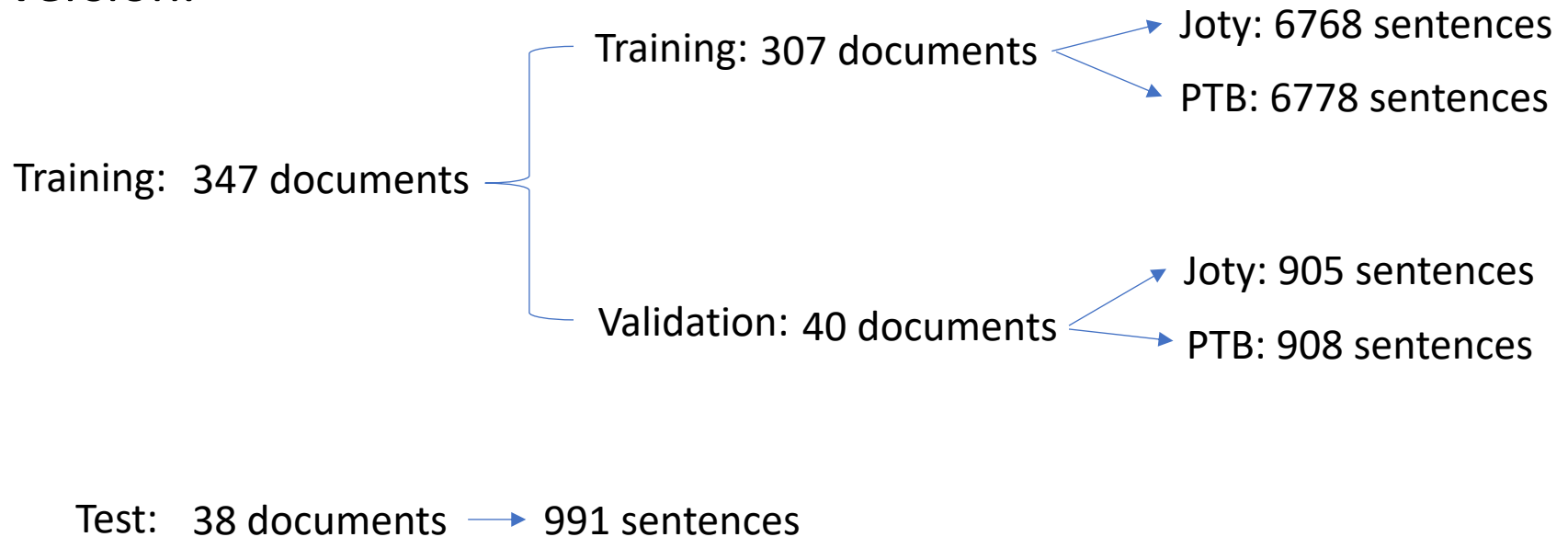
- K candidates are selected from all generated paths according to their CRF scores.



Experiment 1: Discourse segmentation without labels

Dataset: RST-DT <https://catalog.ldc.upenn.edu/products/LDC2002T07>

- Same as [rstfinder](#), we extract 40 validation documents from the original training dataset.
- There are two versions of training and validation dataset, the version of [Joty](#) removes footnote sentences. We utilized Joty's version.



Experiment 1: Discourse segmentation without labels

Parameter setting

- **Model setting**

Model	Pretrained-transformer-MPNet
Optimizer	adam
Learning rate	0.00009
Batch size	8192 tokens
Warm up steps	2.4 epoch
Epoch	30
Attention layer	12
Attention head	12
word embedding size	768
Hidden size	3072
Vocab size	30527 (Bert)
Tokenizer	Byte pair encoder

- **Candidate settings**

Data	Best K path	# of Data
Training data	20	140, 924
Valid / Test data	5	

Experiment 1: Discourse segmentation rerank without labels

Result (*: $p < 0.01$, compare with Baseline_1)

	Precision	Recall	F1
Baseline_1 (published model in their github)	92.22	95.35	93.76
Baseline_2 (average 5 runs, train on joty data)	93.16	96.26	94.68
Baseline_mpnet(average 5 runs, train on joty data)	92.84	95.63	94.21

Table 1. Experiment results of baseline models on the test dataset

		Precision	Recall	F1
Oracle		97.73	98.67	98.20
MPNet Lr0.00009	ensemble 5 runs	83.09	94.98	88.64
	average 5 runs	82.00	94.18	87.68
MPNet Lr0.0002	ensemble 5 runs	85.34	94.91	89.87
	average 5 runs	83.88	94.42	88.83
MPNet + candidate_loss (19+1 candidates)	ensemble 5 runs	95.31	97.56	96.43*
	average 5 runs	95.21	94.46	94.71
+ Concat label embed	ensemble 5 runs	95.05	97.86	96.44*
	average 5 runs	94.86	95.20	94.98
+ Average label embed	ensemble 5 runs	95.54	97.93	96.72*
	average 5 runs	89.04	97.76	92.66

*Learning rate
is too large
for one
random seed*

Table 2. Experiment results of models on the test dataset by given 5 candidates from Baseline_1

Experiment 2: Discourse parsing rerank

Example of input sentence

- 1. For a given sentence, generate the candidate parsing by the baseline model. We utilized format2 in our code.

- With all labels

Format1 : (CONDITION (N (ATtribution (S The Treasury also said)_S (N noncompetitive tenders will be considered timely)_N)_{ATtribution})_N (S (TEMPORAL (N if postmarked no later than Sunday, Oct.29,)_N (N and received no later than tomorrow.)_N)_{TEMPORAL})_S)_{CONDITION}

Format2: (ATtribution (S The Treasury also said)_S)_{ATtribution} (SPAN (N noncompetitive tenders will be considered timely)_N)_{SPAN}

- With nuclearity labels (N: Nucleus , S: Satellite)

(N (S The Treasury also said)_S (N noncompetitive tenders will be considered timely)_N)_N (S (N if postmarked no later than Sunday, Oct.29,)_N (N and received no later than tomorrow.)_N)_S

- With relation labels

Format1 : (CONDITION (ATtribution The Treasury also said | noncompetitive tenders will be considered timely)_{ATtribution} (TEMPORAL if postmarked no later than Sunday, Oct.29, | and received no later than tomorrow.)_{TEMPORAL})_{CONDITION}

Format2: (ATtribution The Treasury also said)_{ATtribution} (SPAN noncompetitive tenders will be considered timely)_{SPAN}

- With span labels

((The Treasury also said) (noncompetitive tenders will be considered timely)) ((if postmarked no later than Sunday, Oct.29,) (and received no later than tomorrow.))

Experiment 2: Discourse parsing rerank

Example of input sentence

- 2. Build input sentence (with nuclearity labels)

(N (S The Treasury also said)_S (N noncompetitive tenders will be considered timely)_N)_N (S (N if postmarked no later than Sunday, Oct.29,)_N (N and received no later than tomorrow.)_N)_S

Replace table:

(S	→	(satellite_left	→	[unused0]
) _S	→	satellite_right)	→	[unused1]
(N	→	(nucleus_left	→	[unused2]
) _N	→	nucleus_right)	→	[unused3]

The vocabulary of MPNet is the same as Bert which contains tokens from [unused0] to [unused993]. These [unused*] tokens are specifically used as randomly initialized embeddings

Replaced sentences:

(nucleus_left (satellite_left The Treasury also said satellite_right) (nucleus_left noncompetitive tenders will be considered timely nucleus_right) nucleus_right) (satellite_left (nucleus_left if postmarked no later than Sunday, Oct.29, nucleus_right) (nucleus_left and received no later than tomorrow. nucleus_right) satellite_right)

[unused2] [unused0]The Treasury also said [unused1] [unused2] noncompetitive tenders will be considered timely [unused3] [unused3] [unused0][unused2] if postmarked no later than Sunday, Oct.29, [unused3] [unused2] and received no later than tomorrow. [unused3] [unused1]

Experiment 2: Discourse parsing rerank

Example of input sentence

- 3. Build input sentence (with relation labels)

19 relations * 2 directions = 38

(CONDITION	→	(condition_left	→	[unused0]
) _{CONDITION}	→	condition_right)	→	[unused1]
(ATtribution	→	(attribution_left	→	[unused2]
) _{ATtribution}	→	attribution_right)	→	[unused3]
...	

label embeddings with its definition

List of label definitions ([Discourse Tagging Reference Manual](#))

Span	Definition
span	span

Nuclearity	Definition
nucleus	a more salient or essential piece of information
satellite	a supporting or background piece of information

Relation	Definition
Attribution	attribution, attribution represents both direct and indirect Instances of reported speech
Background	background or circumstance
Cause	cause or result
Comparison	comparison , preference, analogy or proportion
Condition	condition , hypothetical, contingency or otherwise
contrast	contrast relation, spans contrast with each other along some dimension. Typically, it includes a contrastive discourse cue, such as but, however, while.

label embeddings with its definition

List of label definitions ([Discourse Tagging Reference Manual](#))

Relation	Definition
Elaboration	elaboration, elaboration provides specific information or details to help define a very general concept
Enablement	enablement , enablement presentes action to increase the chances of the unrealized situation being realized.
Evaluation	evaluation, interpretation, conclusion or comment
Explanation	evidence, explanation or reason
Joint	list, list contains some sort of parallel structure or similar fashion between the units
Manner-Means	explaining or specifying a method , mechanism , instrument , channel or conduit for accomplishing some goal

label embeddings with its definition

List of label definitions ([Discourse Tagging Reference Manual](#))

Relation	Definition
Topic-Comment	problem solution , question answer , statement response, topic comment or rhetorical question
Summary	summary or restatement
Temporal	situations with temporal order, before, after or at the same time
Topic change	topic change
textual-organization	links that are marked by schemata labels
span	span
same-unit	links between two non-adjacent parts when separated by an intervening relative clause or parenthetical

Experiment 2: Discourse parsing rerank

Dataset: RST-DT <https://catalog.ldc.upenn.edu/products/LDC2002T07>

- We utilized the preprocessed dataset of previous discourse segmentation task.
- In discourse parsing task, because one sentence should contain at least 2 EDU parts, we filtered these sentences from all sentences.

Training: 307 documents → Joty: 6768 sentences → 4524 sentences

Validation: 40 documents → Joty: 905 sentences → 636 sentences

Test: 38 documents → 991 sentences → 602 sentences

Experiment 2: Discourse parsing rerank

Select baseline model

- We selected [Two-Stage parser](#) as our baseline model. Two-Stage parser (C=1') is trained by Kobayashi-san

	Model	Span (F1)	Nuclearity (F1)	Relation (F1)
Reported by author	Two-Stage parser (Doc-level)	95.6	87.8	77.6
	UnifiedParser(medium)	96.37	89.04	79.03
	UnifiedParser(large) + α	97.44	91.34	81.70
Trained by us	Two-Stage parser (C=1')	97.80	91.76	82.03
	Two-Stage parser (ensemble Svc and Logistic Regression)	97.57	91.60	80.97
	Two-Stage parser (ensemble C=0.5, C=1, C=2, C=3, C=1')	97.53	91.25	81.87
	SpanBasedParser(AAAI20, single)	96.67	90.23	74.76

Table 3. Experiment results of models on the test dataset

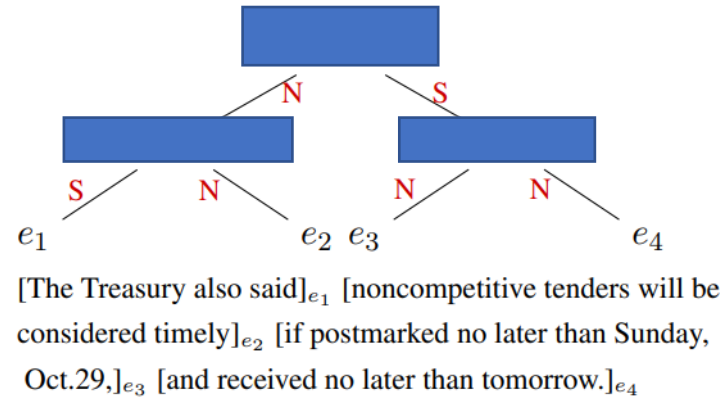
Experiment 2: Discourse parsing rerank

Baseline model: A Two-stage Parsing Method for Text-level Discourse Analysis

- **First stage: Tree Structure Construction**

- Actions:

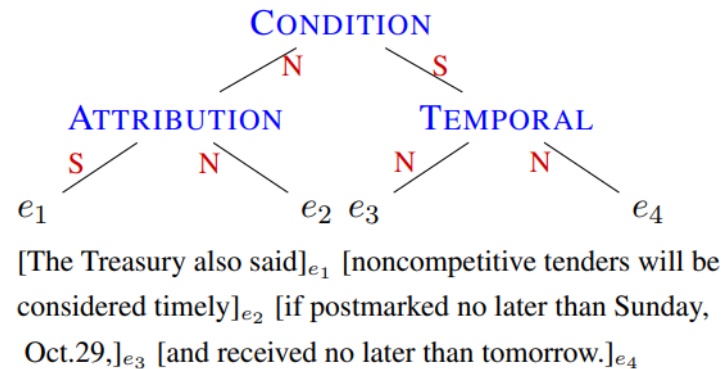
Shift,
Reduce-NN,
Reduce-NS,
Reduce-SN



- **Second stage: Relation Labeling**

- Relations (19 types):

Attribution,
Background,
Cause,
...



Experiment 2: Discourse parsing rerank

Parameter setting

- **Candidate settings**

Data	First stage	Second stage	keep candidates	# of Data
Training data with span/nuclearity labels	20	1	20	60742
Training data with relation/all labels	1	20	20	95004
Valid / Test data	5	5	5	

- **Other parameters are same as previous settings, except for the followings.**

Warm up steps (Span / Nuclearity) 864 (2.4 epoch)

Warm up steps (Relation) 1196 (2.4 epoch)

Experiment 2: Discourse parsing rerank

Result (*: $p < 0.01$, †: $p < 0.05$, compare with Two-Stage parser (C=1', beam search))

Model		Span (F1)	Nuclearity (F1)	Relation (F1)
Oracle		98.70	95.53	90.19
Two-Stage parser (C=1', beam search)		97.92	92.07	82.06
MPNet with Span labels	ensemble 5 models	98.23†	92.31	82.22
	average 5 models	97.97	91.85	81.67
MPNet with nuclearity labels	ensemble 5 models	98.31†	94.00*	83.63*
	average 5 models	98.04	92.49	81.87
MPNet with relation labels	ensemble 5 models	98.00	93.09*	83.99*
	average 5 models	97.75	92.30	82.54
MPNet with all labels	ensemble 5 models	97.84	92.90†	84.11*
	average 5 models	97.76	92.24	81.72
MPNet with Span labels aveemb	ensemble 5 models	98.27†	92.39	82.42
	average 5 models	98.19	92.27	82.14
MPNet with nuclearity labels aveemb	ensemble 5 models	98.31*	93.88*	83.56*
	average 5 models	98.23	93.55	83.27
MPNet with relation labels aveemb	ensemble 5 models	98.12	93.13*	84.69*
	average 5 models	98.12	92.89	83.97
MPNet with all labels aveemb	ensemble 5 models	98.04	92.74	84.18*
	average 5 models	97.79	92.24	81.52

Table 4. Experiment results of models on the test dataset by given 5 candidates from Two-Stage parser