A Language Model-based Generative Classifier for Sentence-level Discourse Parsing

Ying Zhang **Hidetaka Kamigaito** **Manabu Okumura**

{ zhang, kamigaito, oku } @ lr.pi.titech.ac.jp

東京工業大学
Tokyo Institute of Technology

Content

Motivation - Rhetorical Structure Theory

- Rhetorical Structure Theory (RST) is applied to describe an internal discourse structure for the text as a constituent tree (Mann and Thompson, 1988).

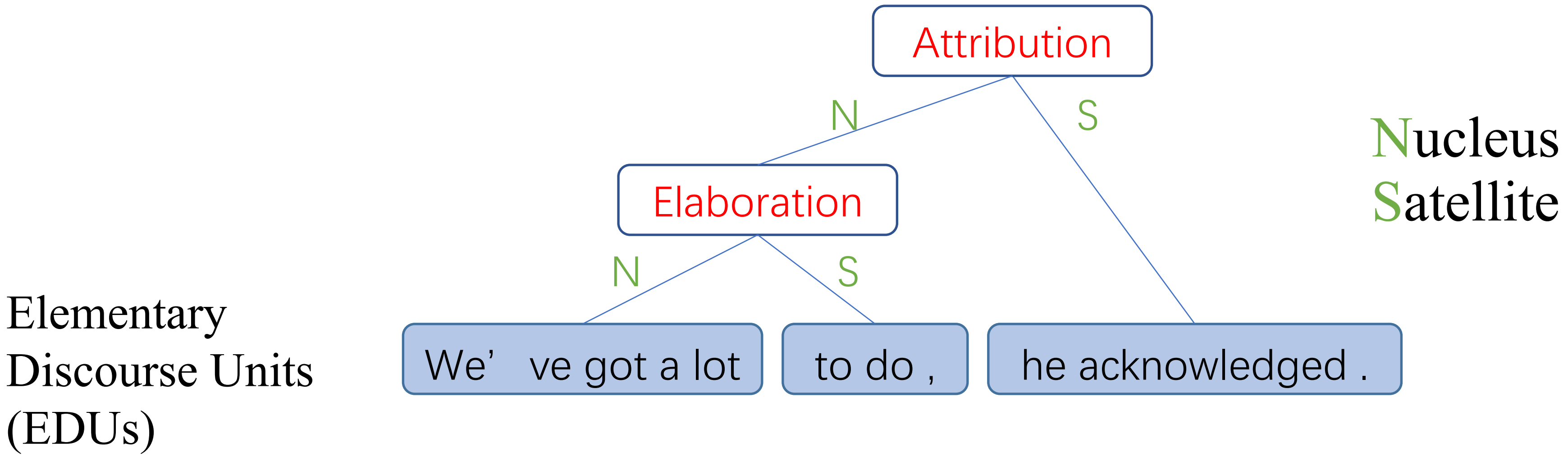


Figure 1. An example RST discourse tree

Motivation - Related Work

- **Discourse segmentation**: a task to detect EDU boundaries in a given text.

- **Discourse parsing**: a task to link spans for detected EDUs.

Most prior work is on the basis of discriminative models $P(\boldsymbol{y}|\boldsymbol{x})$, which learn mapping from input texts $\boldsymbol{x}$ to predicted labels $\boldsymbol{y}$.

The number of labeled RST discourse trees is restricted.

Motivation - Related Work

- Thus, there still remains room for improving model performance by **considering mapping from predictable labels to input texts** to exploit more label information.

In this research, we propose a *language model-based generative classifier* (LMGC) as a reranker for both discourse segmentation and sentence-level discouse parsing.

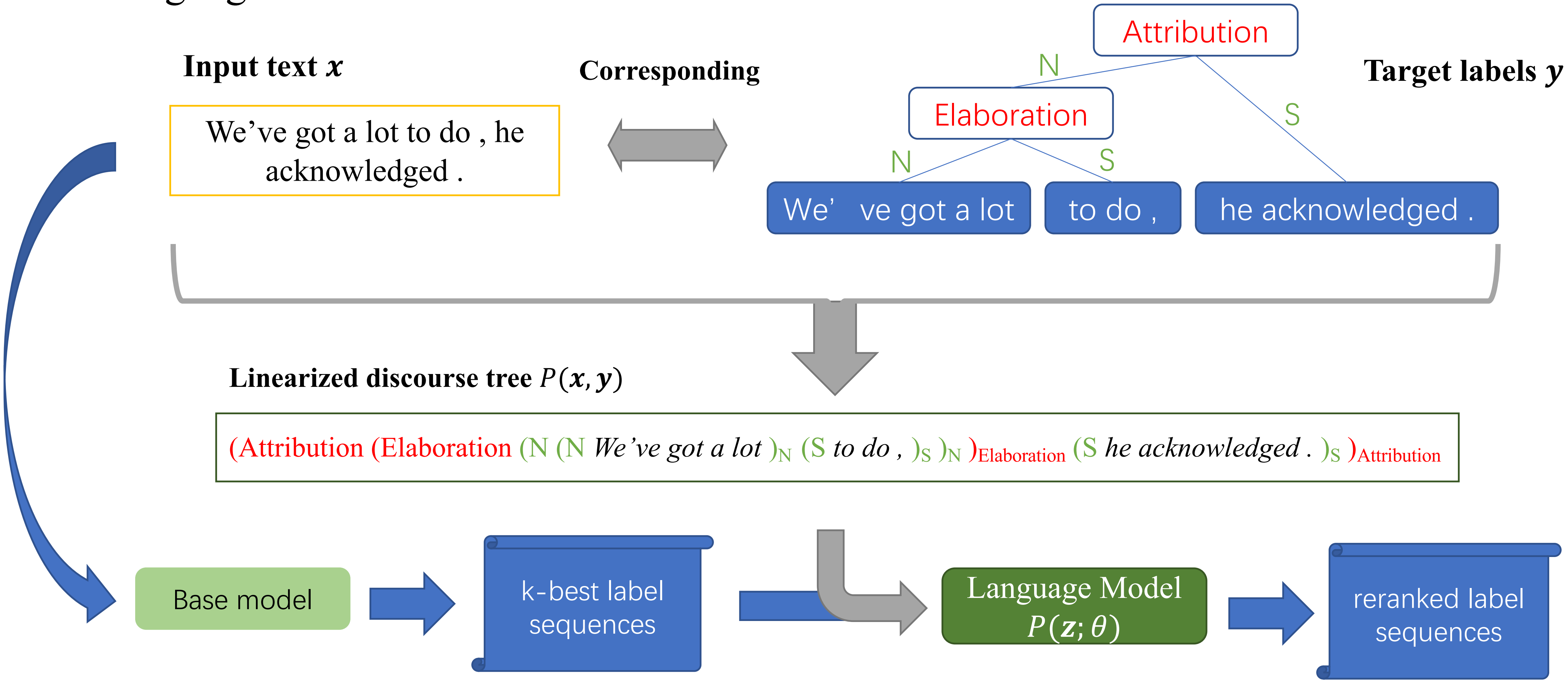
- A BERT-style bidirectional Transformer encoder (Devlin et al., 2019)
- Joint probability $P(\boldsymbol{x}, \boldsymbol{y})$ of an input text and its predictable labels
- Adopt pre-trained language models such as MPNet (Song et al., 2020)

Content

- Motivation
- Language Model-based Generative Classifier (LMGC)
- Experiment
- Results
- Conclusion

LMGC - Overview

- Predicting joint probability $P(\boldsymbol{x}, \boldsymbol{y})$ for a linearized discourse tree based on the language model.



LMGC - Joint Representation

Sentence with EDU boundary labels $P(\boldsymbol{x}, \boldsymbol{e})$

e_1 [EDU] e_2 [EDU] e_3 [EDU]

Sentence with span labels $P(\boldsymbol{x}, \boldsymbol{e}, \boldsymbol{s})$

(Span (Span e_1 $)_{\text{Span}}$ (Span e_2 $)_{\text{Span}}$ $)_{\text{Span}}$ (Span e_3 $)_{\text{Span}}$

Sentence with nuclearity labels $P(\boldsymbol{x}, \boldsymbol{e}, \boldsymbol{u})$

(N (N e_1 $)_{\text{N}}$ (S e_2 $)_{\text{S}}$ $)_{\text{N}}$ (S e_3 $)_{\text{S}}$

Sentence with relation labels $P(\boldsymbol{x}, \boldsymbol{e}, \boldsymbol{r})$

(Span (Span e_1 $)_{\text{Span}}$ (Elaboration e_2 $)_{\text{Elaboration}}$ $)_{\text{Span}}$ (Attribution e_3 $)_{\text{Attribution}}$

e_i represents the corresponding EDU for sentence *We' ve got a lot to do, he acknowledged.*

LMGC - Joint Probabilities

- Follow the decomposition of pseudo-log-likelihood scores (PLL) (Salazar et al., 2020) , we decompose and calculate logarithmic $P(\boldsymbol{z})$ for sequence $\boldsymbol{z} = (z_1, \dots, z_a)$ as follows:

M_t: **[MASK]** at step t

$$\log P(\boldsymbol{x}, \boldsymbol{y}) = \log P(\boldsymbol{z}; \theta) \approx PLL(\boldsymbol{z}; \theta) \approx \sum_{t=1}^{a} \log P(z_t | z_{<t}, z_{>t}, M_t; \theta)$$

M_t: **[MASK]**

Joint sequence $\boldsymbol{z}$: ? *'ve got a lot* **[EDU]** *to do ,* **[EDU]** *he acknowledged .* **[EDU]**

We choose pretrained masked and permuted language modeling (MPNet) (Song et al., 2020) as our language model and $P(z_t | z_{<t}, z_{>t}, M_t; \theta)$ is computed by two−stream self−attention (Yang et al., 2019).

LMGC - Label Embedding

Joint sequence ***z***: *We've got a lot* **[EDU]** *to do ,* **[EDU]** *he acknowledged .* **[EDU]**

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

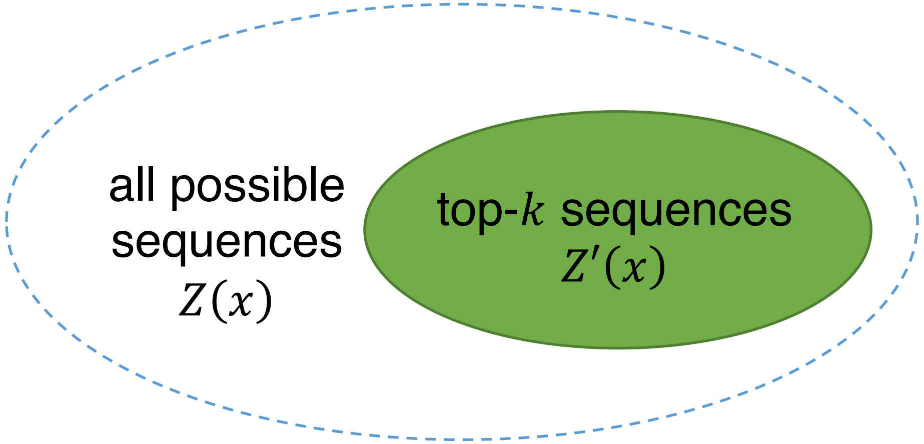
- LMGC – Enhance :

Embedding of **[EDU]** = Ave (Embedding of elementary,
Embedding of discourse,
… ,
Embedding of tree)

- LMGC – Extend :

Joint sequence ***z***: *We've got a lot* **[EDU]** *to do ,* **[EDU]** *he acknowledged .* **[EDU]** [EDU] :elementary discourse units are the minimal building blocks of a discourse tree

LMGC – Objective function



$Z'(x)$ is generated by a base model.

We denote $z_g \in Z(x)$ as the correct joint sequence of x and assume that O_a lists all permutations of set $\{1, 2, \dots, a\}$.

For $z \in Z'(x) \cup \{z_g\}$, we train the model parameter θ in LMGC by maximizing the following expectation over all permutations:

$$\mathbb{E}_{o \in O_a} \sum_{t=c+1}^{a} [I_z \log P\left(z_{o_t} \middle| z_{o_{<t}}, M_{o_{>c}}; \theta\right) + (1 - I_z) \log(1 - P\left(z_{o_t} \middle| z_{o_{<t}}, M_{o_{>c}}; \theta\right))]$$

where I_z is the indicator function, defined as follows:

$$I_z := \begin{cases} 1 & if\ z = z_g, \\ 0 & if\ z \neq z_g \end{cases}$$

c : the number of non-predicted tokens $z_{o_{\leq c}}$. $M_{o_{>c}}$: the mask tokens [MASK] at position $o_{>c}$.

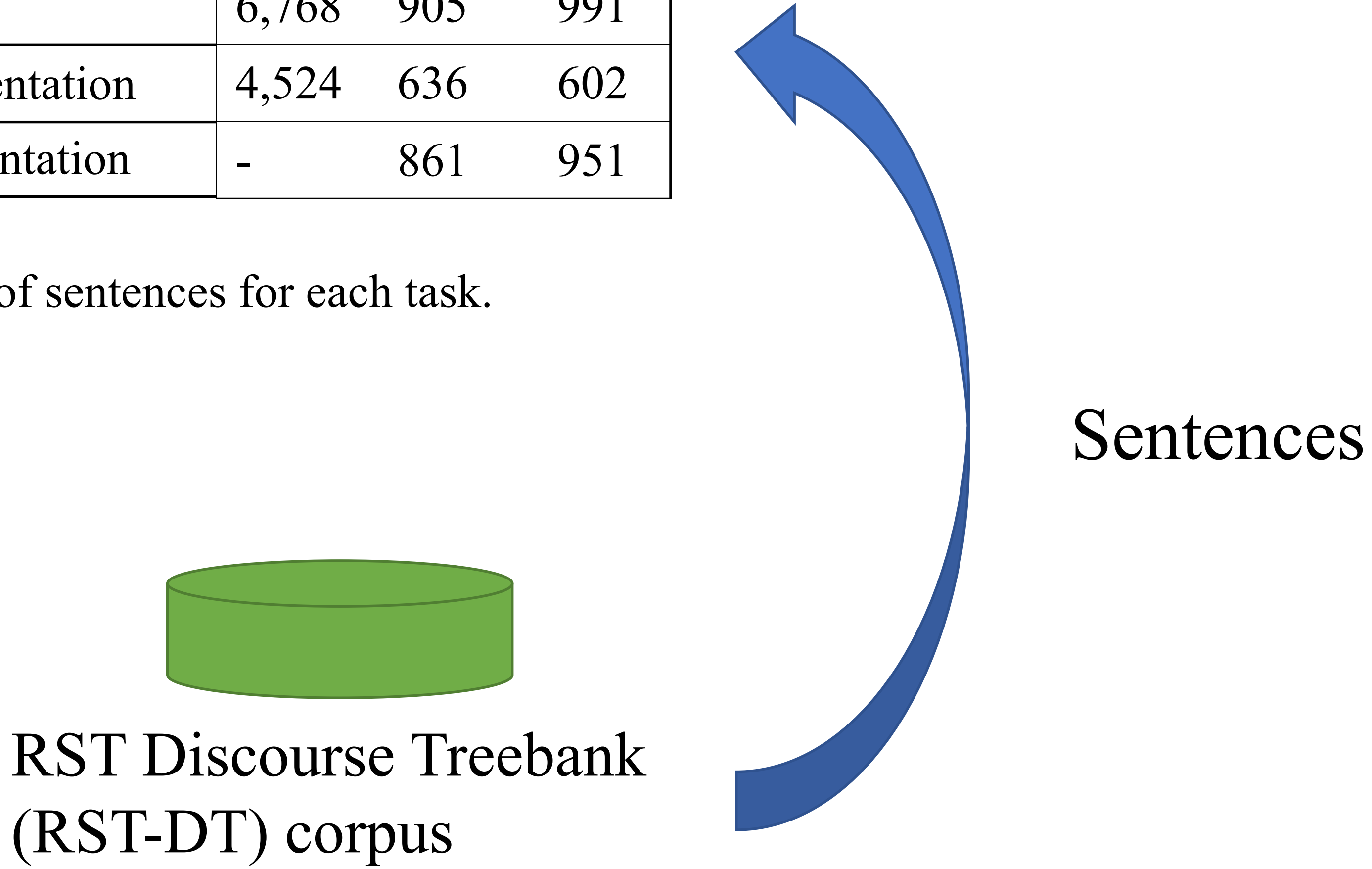
Content

- Motivation
- Language Model-based Generative Classifier (LMGC)
- Experiment
- Results
- Conclusion and Future Work

Experiment - Dataset

Task	Train	Valid	Test
(a) Segmentation	6,768	905	991
(b) Parsing w/ gold segmentation	4,524	636	602
(c) Parsing w/ auto segmentation	-	861	951

Table 1: The number of sentences for each task.



Experiment - Evaluation Metric

(a) Segmentation

Micro-averaged precision, recall, F_1 score for EDUs.

(b) Parsing w/ gold segmentation

Micro-averaged F_1 score for span, nuclearity and relation labels.

(c) Parsing w/ auto segmentation

Micro-averaged F_1 score for EDUs, span, nuclearity and relation labels.

- Significance test: paired bootstrap resampling.

Experiment - Settings

• Pretraind language model	MPNet (Song et al., 2020)
• Base segmenter	BiLSTM-CRF (Wang et al., 2018b)
• Base parser	2-stage Parser (Wang et al., 2017)
• Compared model	GPTLM (GPT2-based language model generative classifier)
• Tuned top-*k* (training)	20
• Tuned top-*k* (prediction)	5

Content

- Motivation
- Language Model-based Generative Classifier (LMGC)
- Experiment
- Results
- Conclusion

Results - (a) Segmentation

Model	**Precision**	**Recall**	$\boldsymbol{F_1}$
Oracle	97.73	98.67	98.20
Pointer-networks*	93.34	97.88	95.55
Base segmenter	92.22	95.35	93.76
$\mathrm{GPT2LM}_e$	94.05	95.72	94.88
LMGC_e	95.31	97.56	96.43†
$\mathrm{Enhance}_e$	**95.54**	**97.93**	**96.72†**
Extend_e	95.05	97.86	96.44†

Table 2: Results for the discourse segmentation task.

* : reported score by Lin et al. (2019). † :the score is significantly superior to GPT2LM with a p-value < 0.01.

Results - (b) Parsing w/ gold segmentation

Model	Span	Nuclearity	Relation
Oracle	98.67	95.88	90.07
Pointer-networks*	97.44	91.34	81.70
Base parser	97.92	92.07	82.06
GPT2LM_r	96.35	88.11	77.86
LMGC_s	98.23‡	92.31	82.22
Enhance_s	98.27‡	92.39	82.42
LMGC_u	**98.31**‡	**94.00**†	83.63†
Enhance_u	**98.31**†	93.88†	83.56†
LMGC_r	98.00	93.09†	83.99†
Enhance_r	98.12	93.13†	**84.69**†

Table 3: Results for the sentence-level discourse parsing task with gold segmentation.

† , ‡ : the score is significantly superior to the base parser with a p-value < 0.01 and < 0.05, respectively.

Results - (c) Parsing w/ auto segmentation

Model	Seg	Parse		
		Span	**Nuclearity**	**Relation**
Pointer-networks*	-	91.75	86.38	77.52
Oracle_{seg}	98.24	-	-	-
Base segmenter	93.92	-	-	-
GPT2LM_e	95.03	-	-	-
LMGC_e	96.51	-	-	-
Enhance_e	**96.79**	-	-	-
Extend_e	96.48	-	-	-
Oracle	-	93.95	91.25	85.93
Base parser	-	93.53	88.08	78.75
GPT2LM_r	-	92.02	84.20	74.49
LMGC_s	-	93.96‡	88.46	79.25
Enhance_s	-	**94.00**†	88.50	79.33
LMGC_u	-	93.96†	**89.90**†	80.33†
Enhance_u	-	93.92‡	89.74†	80.22†
LMGC_r	-	93.65	89.08†	80.57†
Enhance_r	-	93.73	89.16†	**81.18**†

Table 4: Results for the sentence-level discourse parsing task with automatic segmentation.

Content

- Motivation
- Language Model-based Generative Classifier (LMGC)
- Experiment
- Results
- Conclusion

Conclusion

- As a reranker, LMGC achieved the state-of-the-art performances in both discourse segmentation and sentence-level discourse parsing.

- The experimental results showed the potential of constructing label embeddings from token embeddings by using label descriptions.