END-TO-END ARGUMENT MINING FOR DISCUSSION THREADS BASED ON PARALLEL CONSTRAINED POINTER ARCHITECTURE

Tokyo University of Agriculture and Technology, Japan.

Gaku Morio (Master course 2nd)

Katsuhide Fujita (Supervisor)

ArgMining 2018 @ EMNLP 2018

BACKGROUND AND MOTIVATION

Background

- Over the past dozen years or so, middle or large scale online discussions are available through online forums.
  - Recently, **online civic discussions** are also highlighted through the forum [Ito 2014, Park2018].

Takayuki Ito, Yuma Imi, Takanori Ito, and Eizo Hideshima. Collagree: A faciliator-mediated large-scale consensus support system. In Proceedings of the 2nd **International Conference of Collective Intelligence**, 2014.
Joonsuk Park and Claire Cardie. 2018. A corpus of erulemaking user comments for measuring evaluability of arguments. In Proceedings of the Eleventh International Conference on **LREC**, 2018.

The problem is “massive posts.”

- While we **can acquire a lot of posts** in a short time by using the online forum, it is hard to understand all of the posts.
- For example, in the online civic discussion in our previous work [Morio 2018] included,
  - Several days for the discussion;
  - **800+ citizens** who joined the discussion,
  - **1,300+ posts**.

- So, how to understand the enormous *opinions*?
  - We estimate **Argument Mining will do**!

Gaku Morio and Katsuhide Fujita. Predicting argumentative influence probabilities in large-scale online civic engagement. In Companion Proceedings of **The Web Conference 2018**, **WWW ’18**, pp. 1427–1434.

Motivation

- In the present study, we focus on **argument mining** to understand fine-grained opinions in the discussion forum,
  - because extracting **premises** behind citizens' claim is important to understand their ideas.

CONTRIBUTIONS OF OUR WORK

Research Overview

Overview of the contributions

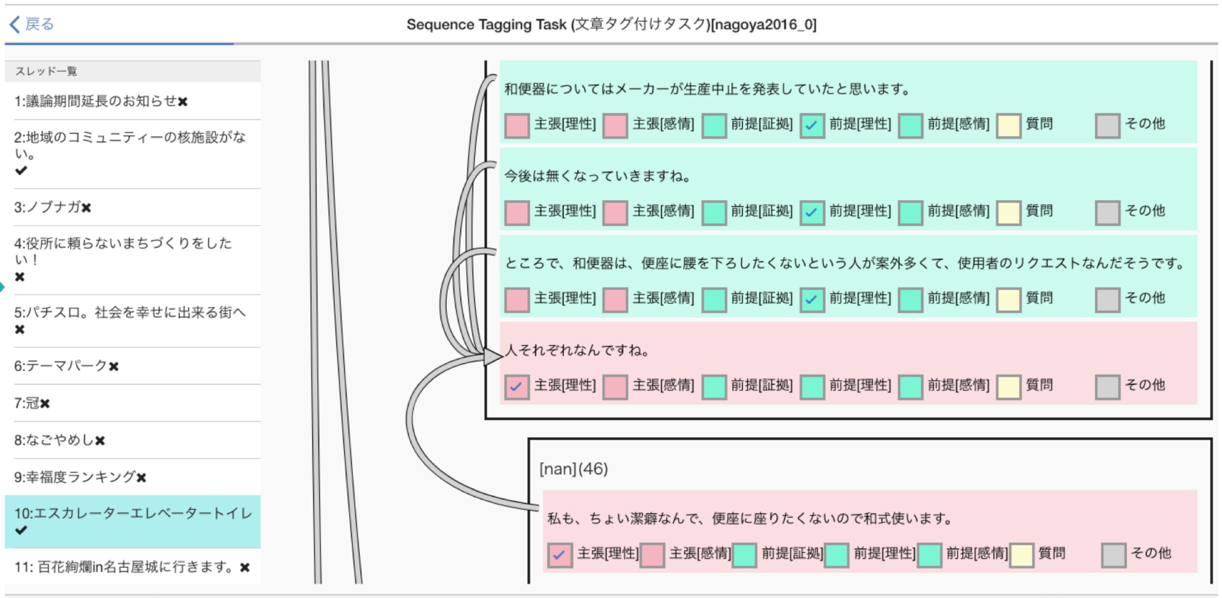
- We tackle **"end-to-end" Argument Mining for discussion forums.**
  - Because there's no definitive studies about it.
  - We provide following **two contributions**;

1. A novel **inner-** and **inter- post scheme**, and **annotations** for discussion threads.
2. **End-to-end classification** approaches for the scheme.
   - **The biggest contribution in this study!**

Contribution overview 1

- **Annotation study for discussion threads.**
  - For this, we provide micro-level **inner-** and **inter-post scheme**.
  - We first conducted the annotation for **Japanese online civic discussion** threads.

Our original annotation tool.

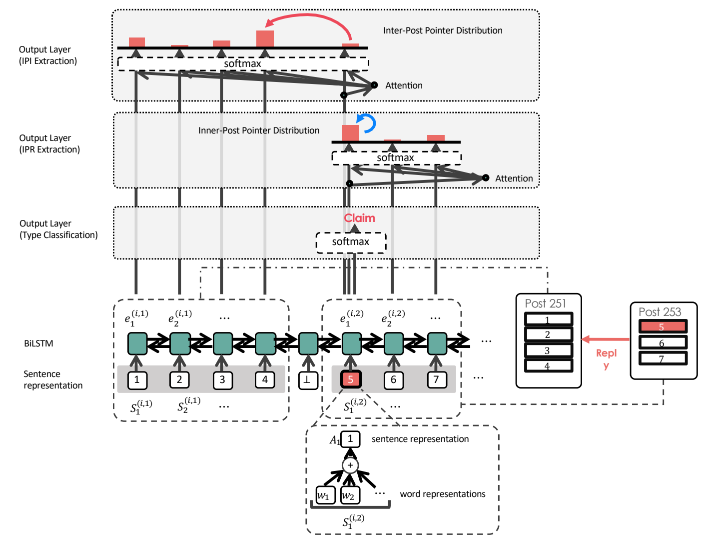


Contribution overview 2

- **Parallel Constrained Pointer Architecture (PCPA)**
  - **PCPA is** a novel **end-to-end neural model** using **Pointer Networks** [Potash 2017].
  - **PCPA can** discriminate;
    - A **sentence type** (i.e., claim, premise or none)
    - An **inner-post relation**;
    - An **inter-post interaction**;

  simultaneously.

Our neural model, PCPA.



- P. Potash, A. Romanov, and A. Rumshisky, “Here’s my point: Joint pointer architecture for argument mining,” in Proceedings of the 2017 Conference on **EMNLP**, 2017.

CONTRIBUTION 1

Annotation Study

Argument Mining for discussion threads

- Related works:
  - There are a few studies which employ micro-level scheme for the discussion thread.
  - Also, most of existing work don’t consider multiple writers in the discussion thread.
    - Though [Hidey 2017] provided a micro-level annotation for the discussion thread, the work don’t distinguish inner- and inter- post scheme.

C. Hidey, E. Musi, A. Hwang, S. Muresan, and K. McKeown, “Analyzing the semantic types of claims and premises in an online persuasive forum,” in Proceedings of the 4th **Workshop on Argument Mining**. 2017, pp. 11–21.

Our scheme for **inner-** post argument

- **We assume** each post as a **stand-alone** discourse.
- Therefore, for each post, an independent argument can be created.

i.e., **claim** and **premise** argument [Stab 2017]

Post:171 *Depth = 1*

Yes, I think making the subway operating 24 hours is appealing.

I want to enjoy Nagoya until late at night.

Inner-post relation (IPR)

Claim

Premise

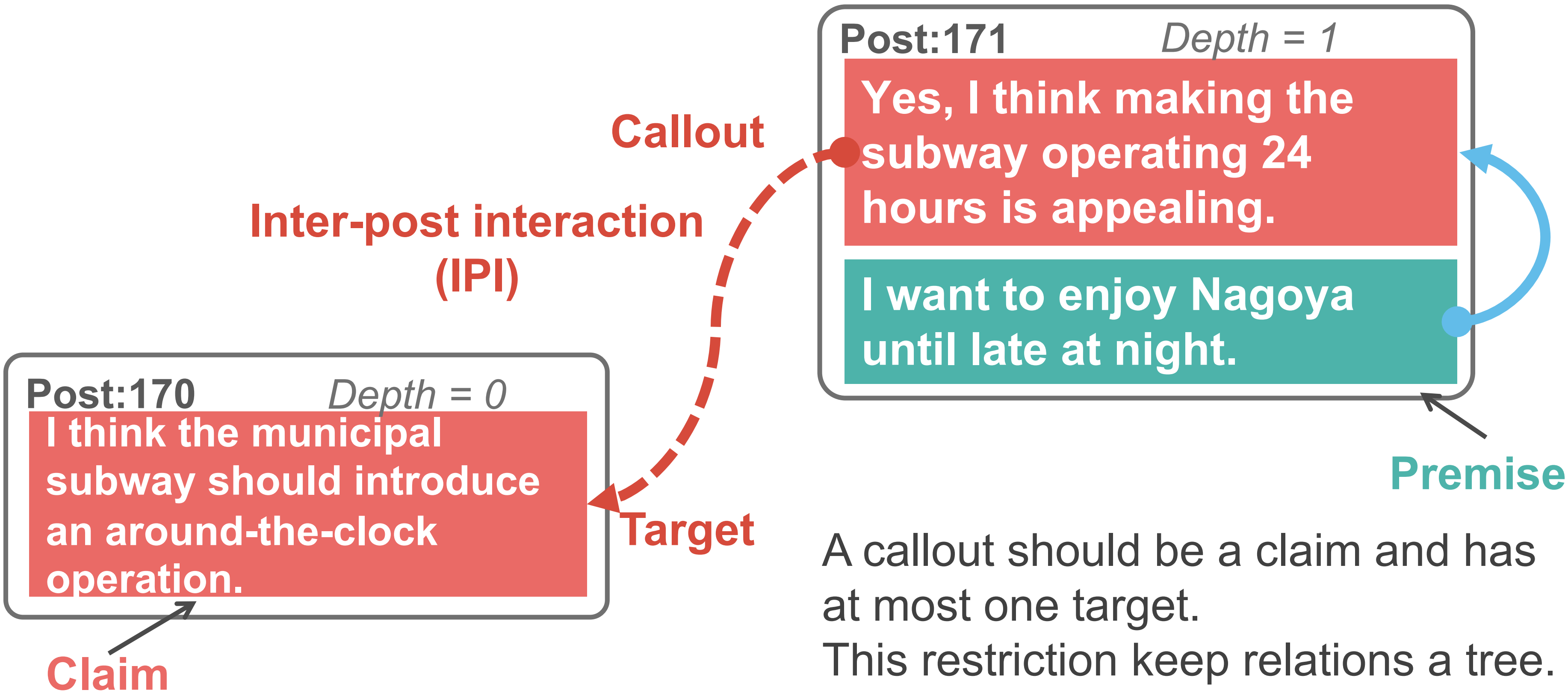
Post:170 *Depth = 0*

I think the municipal subway should introduce an around-the-clock operation.

C. Stab and I. Gurevych, "Parsing argumentation structures in persuasive essays," **Computational Linguistics**, vol. 43, no. 3, pp. 619–659, 2017.

Our scheme for **inter-** post interaction

- **To extract** the **inter-post interaction**, we introduce the interaction model similar to [Ghosh 2014].



A callout should be a claim and has at most one target.
This restriction keep relations a tree.

D. Ghosh, S. Muresan, N. Wacholder, M. Aakhus, and M. Mitsui, "Analyzing argumentative discourse units in online interactions," in Proceedings of the First **Workshop on Argument Mining**, 2014, pp. 39–48.

Annotation

- We annotated our **original online civic discussion**.
  - The online civic engagement was held in **Nagoya city**, Japan, in cooperation with the local government.
  - In this study, we employ "**sentence-level**" annotation because a proposition appears per sentence in most cases.

- The data includes;
  - 399 threads;
  - 1327 posts;
  - 5559 sentences.

Annotation results

- **We acquired state-of-the-art size** of discussion dataset.
  - Also, some properties like a large proportion of premises compared to claims are confirmed.
- However, **inter-annotator agreements are lower** than the essays.
  - We attribute this as following two factors;

1 Most of citizen's comments are not well written.

2 Our sentence-level annotation, rather than token-level.

Corpus	Type	Size	κ
COLLAGREE [ours]	Claim	1449	.531
	Premise	2762	.554
	NonArg	1348	.529
	IPR w/ A0	2762	.466
	IPI	745	.430
Persuasive Essays [Stab2017]	Claim	1506	.635
	Premise	3832	.833
	Inner-essay rel	3832	.708-.737

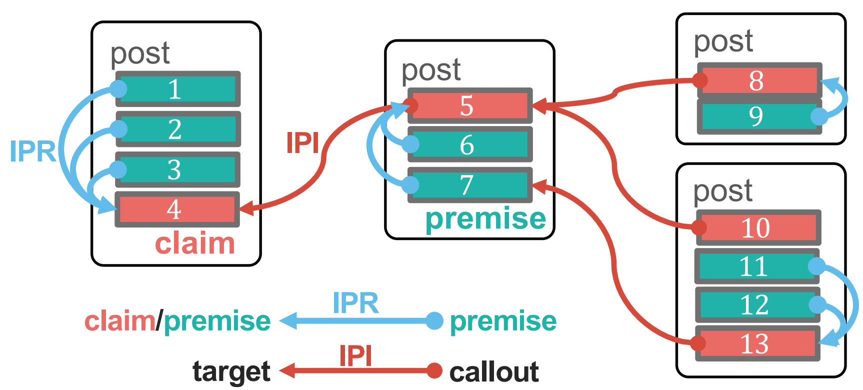
CONTRIBUTION 2

Parallel Constrained Pointer Architecture (PCPA)

Parallel Constrained Pointer Architecture (PCPA)

- **PCPA** is a novel neural model which **can discriminate;**
  - **Claim**;
  - **Premise**;
  - **Inner-post relation** (**IPR**);
  - **inter-post interaction** (**IPI**);

simultaneously (i.e., end-to-end model).



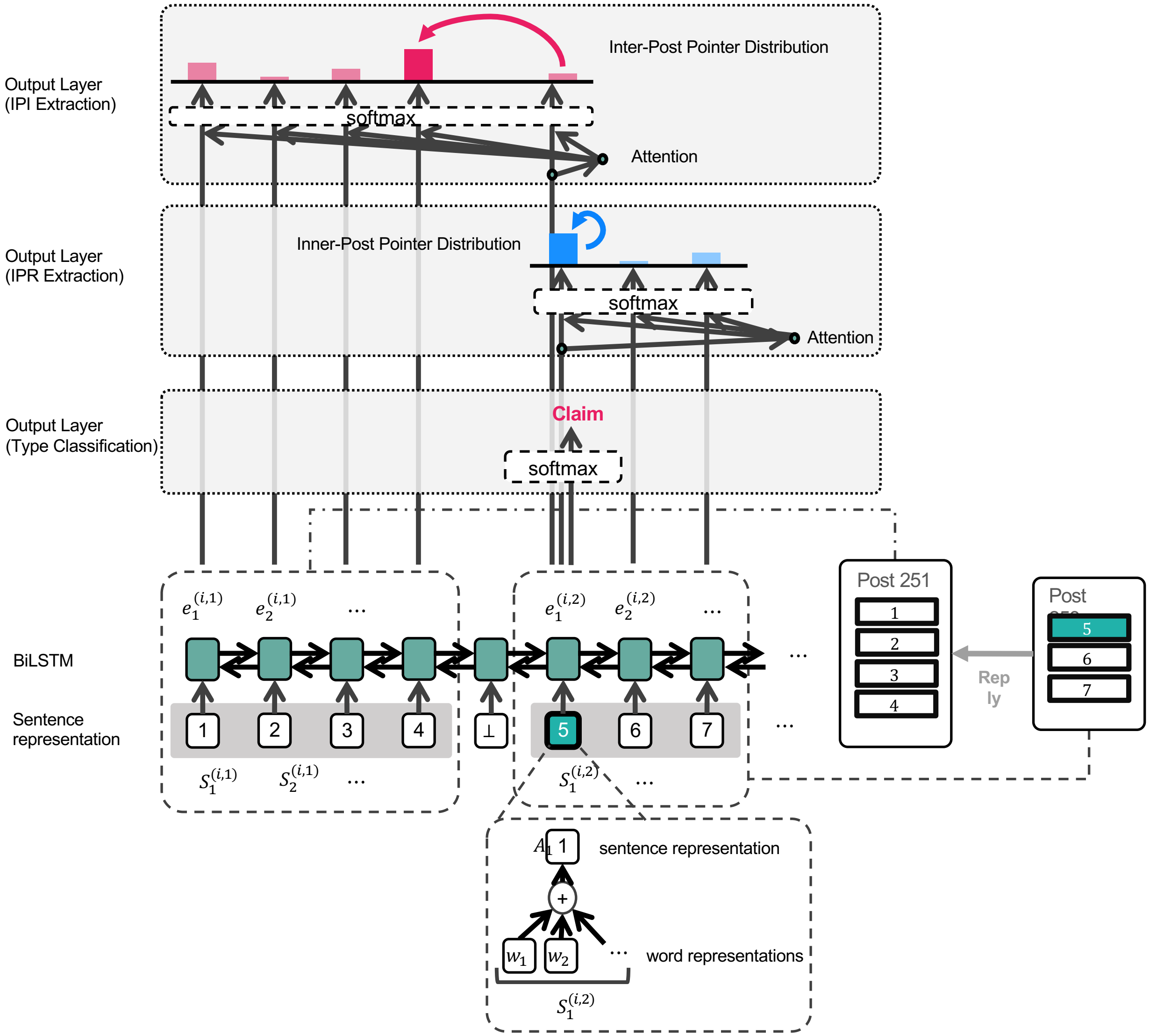
Parallel Constrained Pointer Architecture (PCPA)

- In related works,
  - [Eger 2017] pointed out that end-to-end neural models have advantages in terms of “low error propagation.”
  - Also, [Potash 2017] employed Pointer Networks to discriminate relation target in arguments.
- Thus, in this study we propose an end-to-end model based on Pointer Networks, **PCPA**.
  - Our PCPA has two Pointer Networks for inner- and inter- relation i.e., **parallel** architecture.
  - Our PCPA can effectively constrain computation space based on explicit constraints of discussion threads i.e., **constrained pointer** architecture.
    - So we call our model **Parallel Constrained Pointer Architecture (PCPA)**.

- S. Eger, J. Daxenberger, and I. Gurevych, “Neural end-to-end learning for computational argumentation mining,” in Proceedings of the 55th Annual Meeting of the **ACL**, 2017.
- P. Potash, A. Romanov, and A. Rumshisky, “Here’s my point: Joint pointer architecture for argument mining,” in Proceedings of the 2017 Conference on **EMNLP**, 2017.

PCPA is composed of:

1. Input module
2. Encoding module
3. Output modules

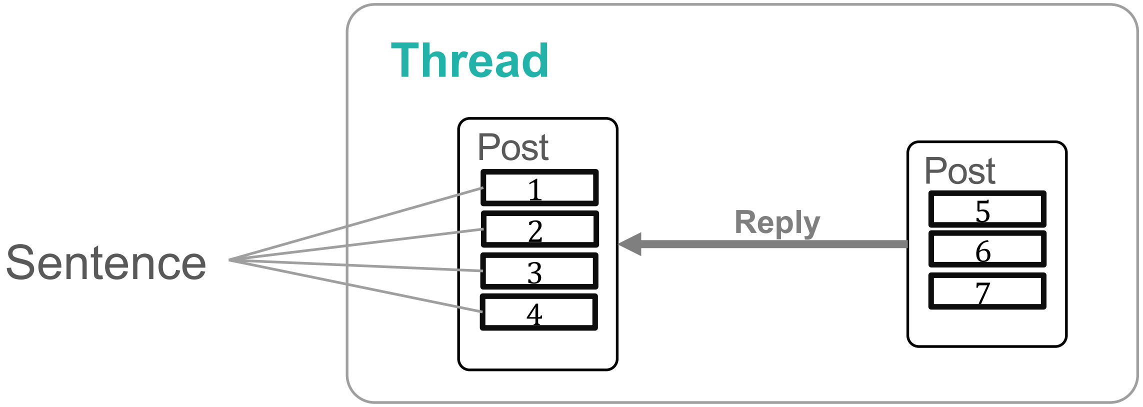


PCPA is composed of:

1. **Input module**
2. Encoding module
3. Output modules

e.g.

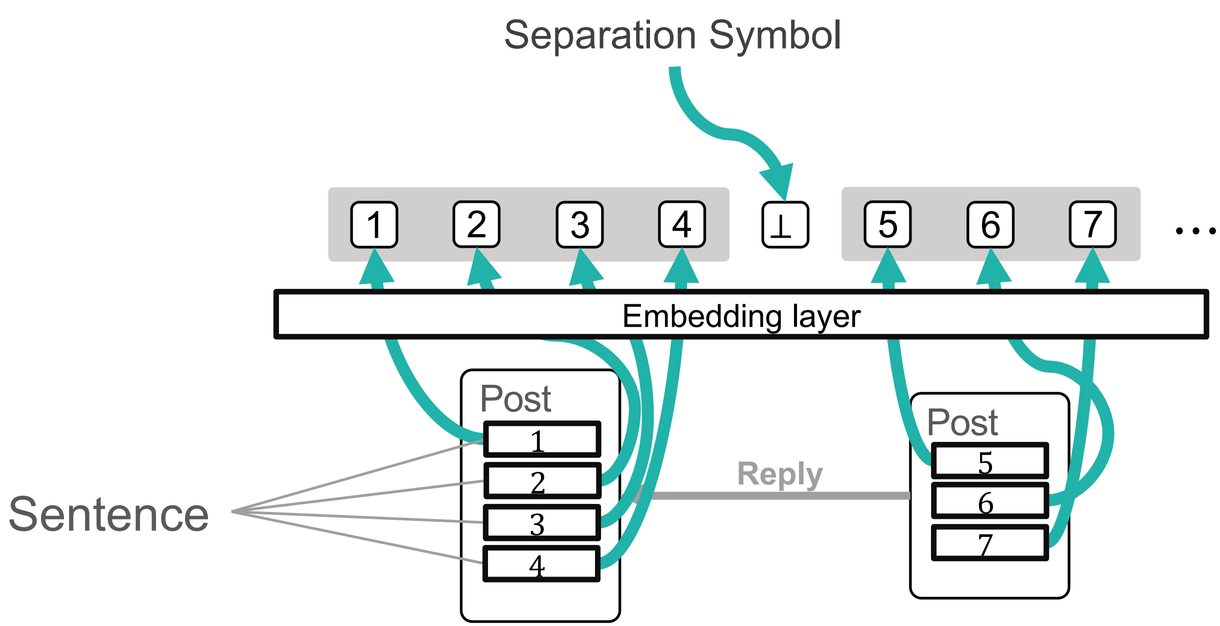
For example, assume given following thread with two posts.



PCPA is composed of:

1. **Input module**
2. Encoding module
3. Output modules

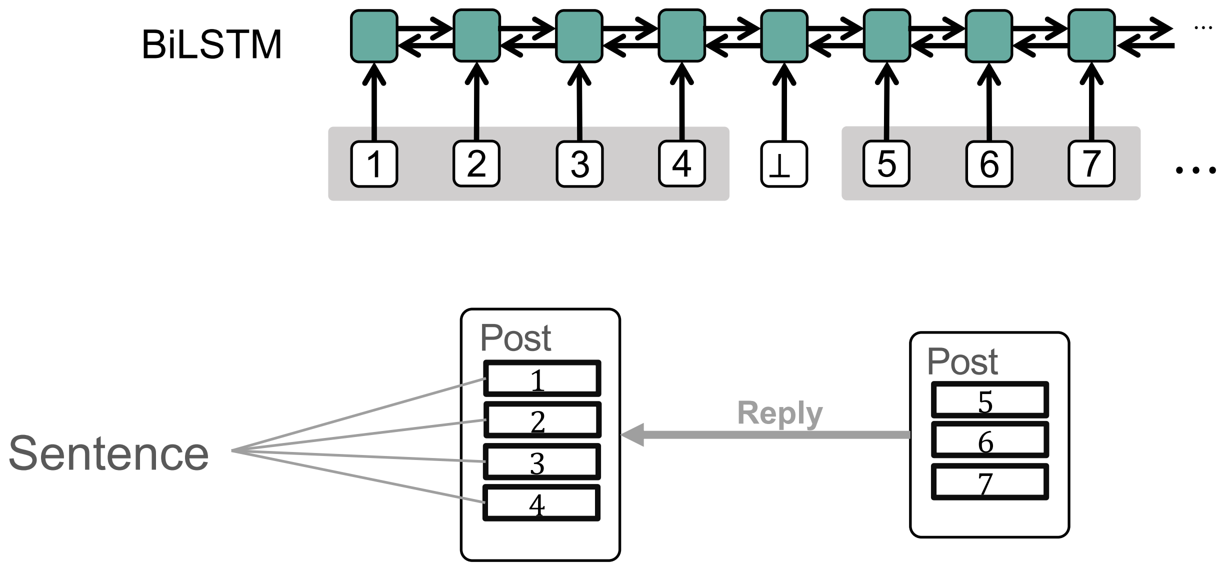
In the input module, each sentence is **converted into sentence representation**.



PCPA is composed of:

1. Input module
2. **Encoding module**
3. Output modules

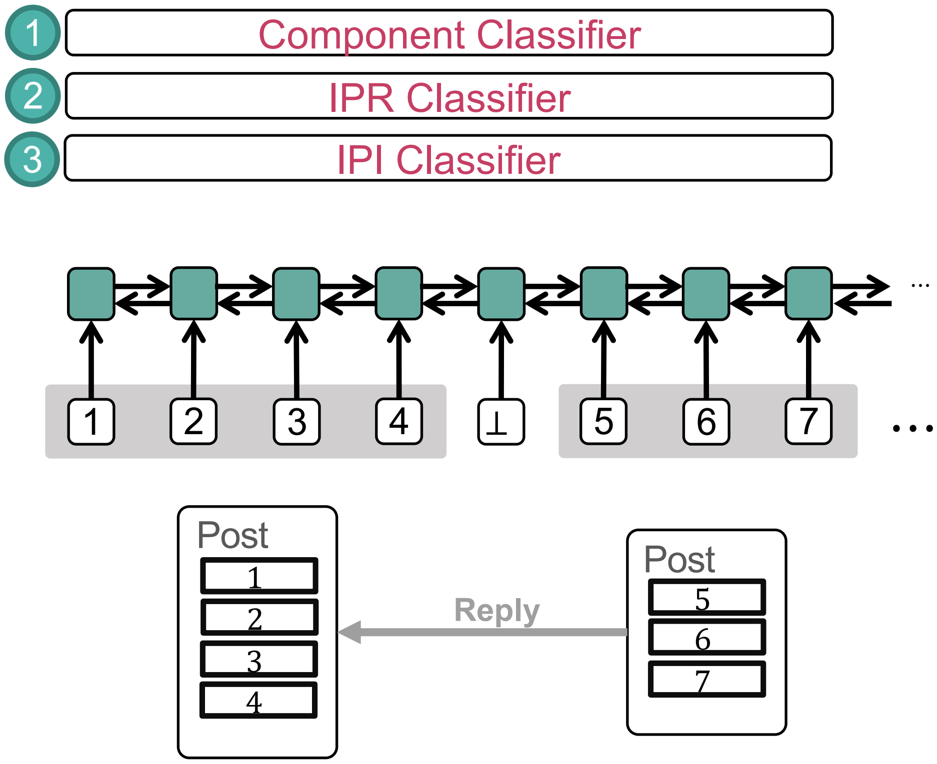
Next, the encoding module with **BiLSTM** acquires context-aware sentence representations.



PCPA is composed of:

1. Input module
2. Encoding module
3. **Output modules**

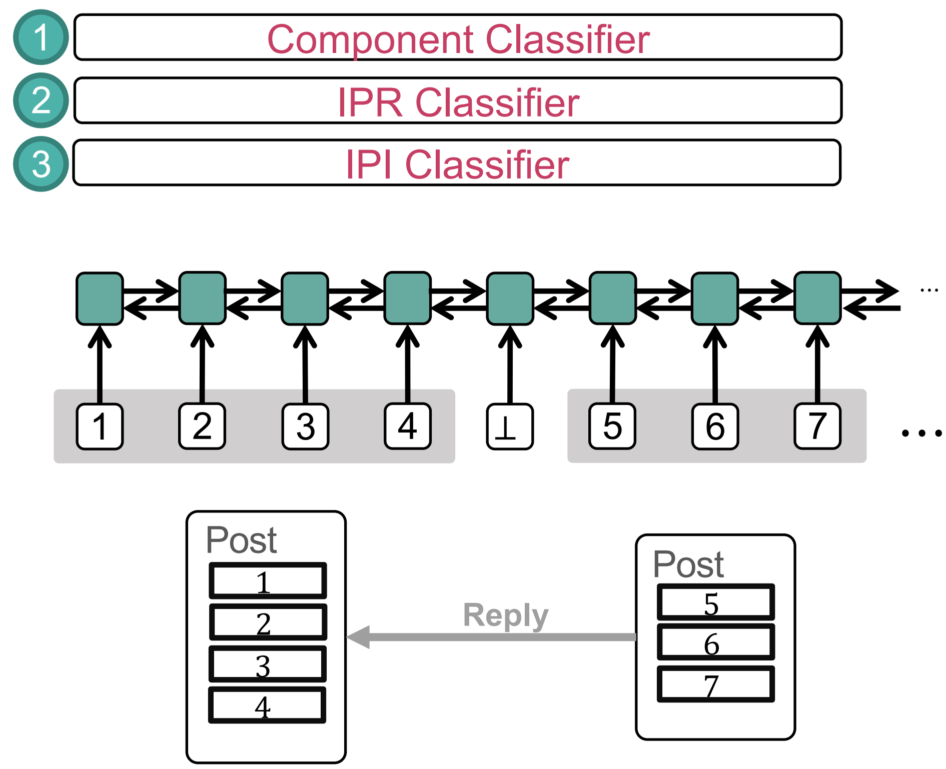
The output modules are PCPA's classification module which has **three output classification layers**.



PCPA is composed of:

1. Input module
2. Encoding module
3. **Output modules**

First, we explain the **Component Classifier**.



PCPA is composed of:

1. Input module
2. Encoding module
3. **Output modules**

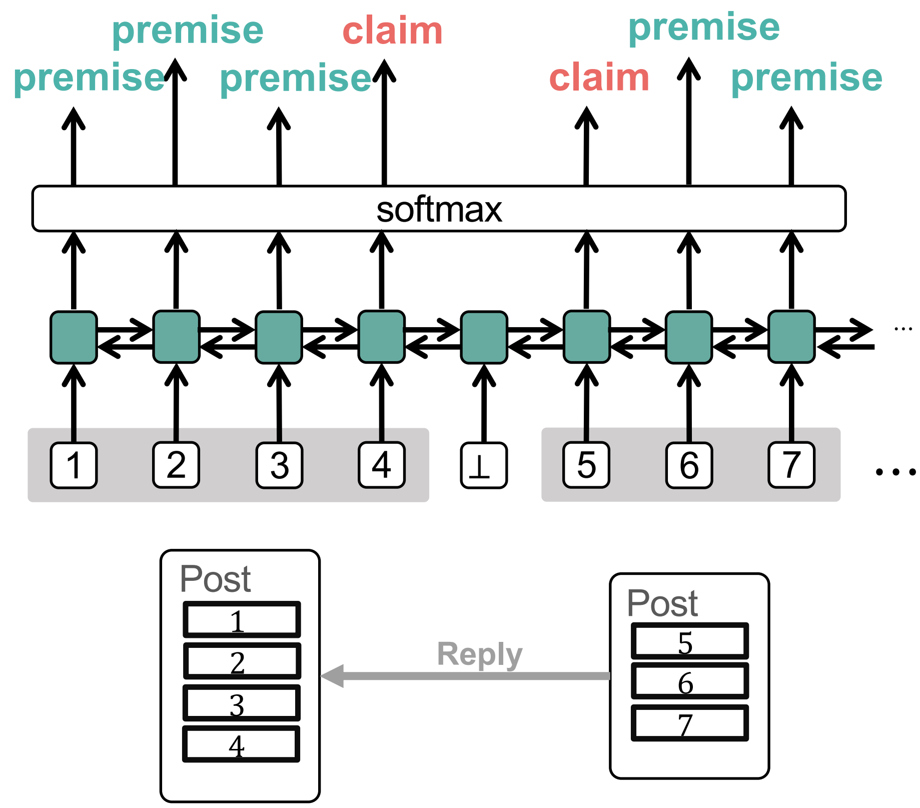
$$p(y_k^{type} \mid P_j^{(i)}) = \mathrm{softmax}(z_k^{(i,j)})$$

$$L_i^{type} = \sum_{j=1}^{N_i} \sum_{k=1}^{N_{i,j}} \log p(y_k^{type} \mid P_j^{(i)})$$

Objective

1 Component Classifier

This layer classifies a **sentence type** (**premise**, **claim** or non-argumentative.)

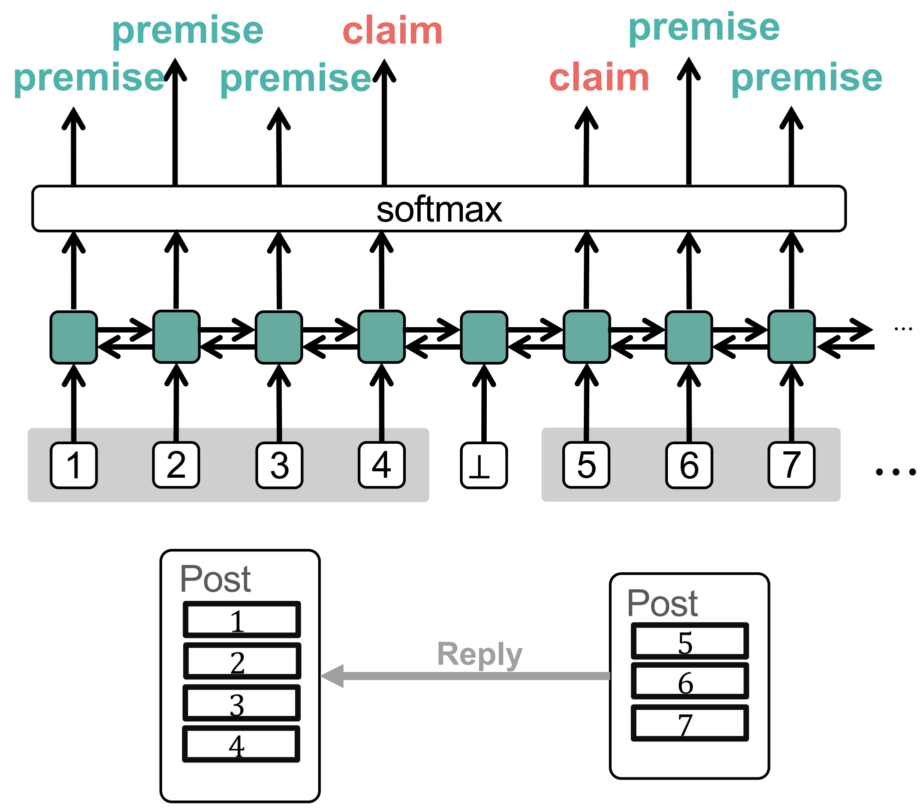


PCPA is composed of:

1. Input module
2. Encoding module
3. **Output modules**

1 Component Classifier

This layer classifies a **sentence type** (**premise**, **claim** or non-argumentative.)



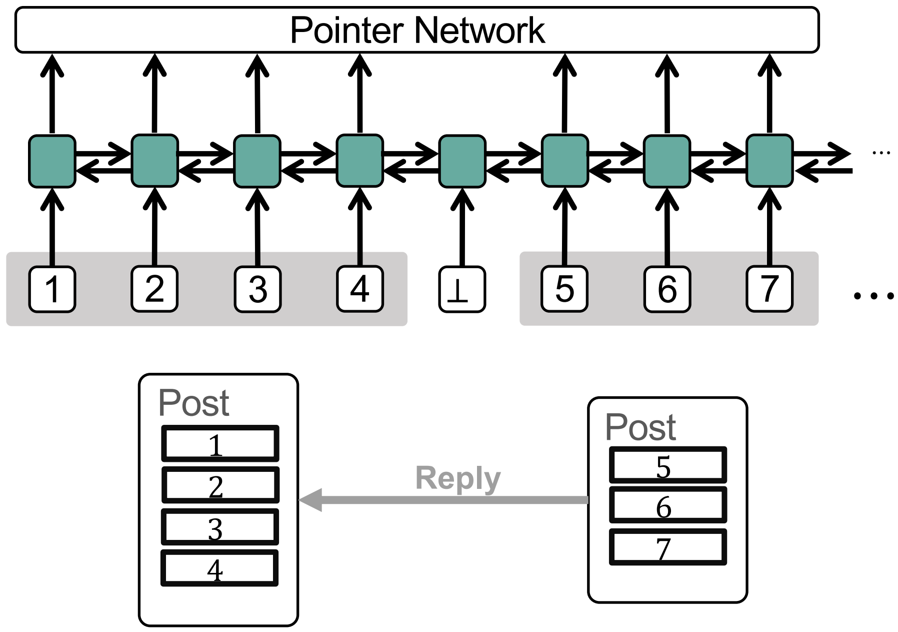
PCPA is composed of:

1. Input module
2. Encoding module
3. **Output modules**

2 IPR Classifier

Next, the IPR Classifier discriminates **inner-post relations** using Pointer Networks.

Pointer Network can estimate the relation target by a pointer distribution.

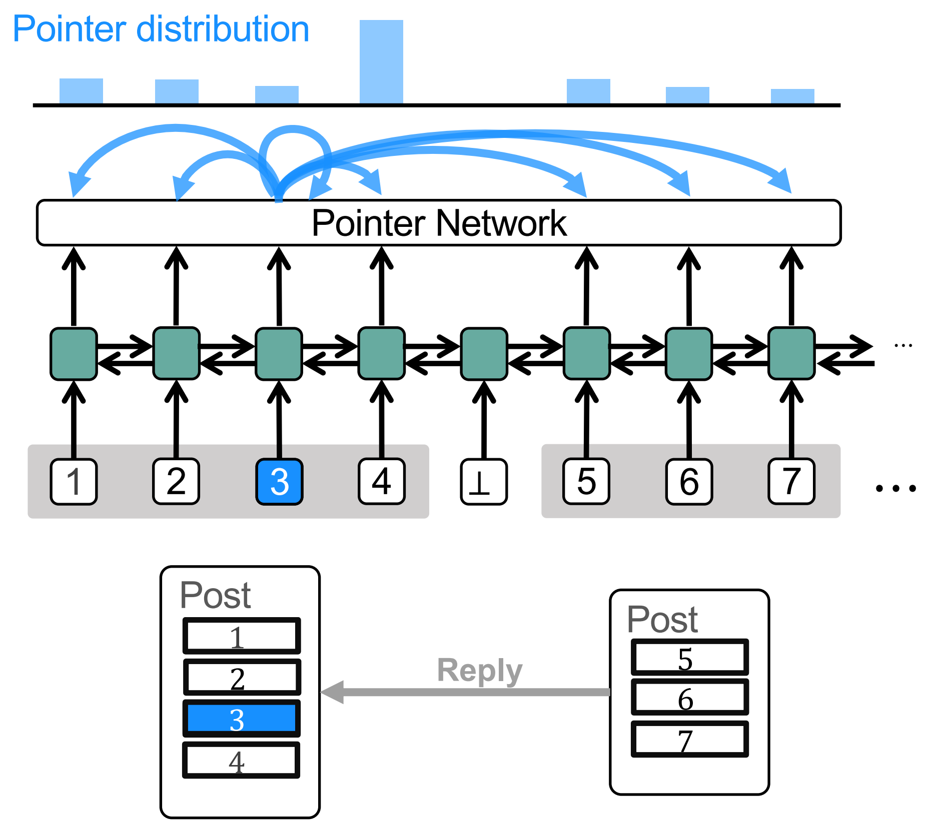


PCPA is composed of:

1. Input module
2. Encoding module
3. **Output modules**

e.g.

For example, let me explain how to search an inner-post relation (IPR) target of sentence “3.”

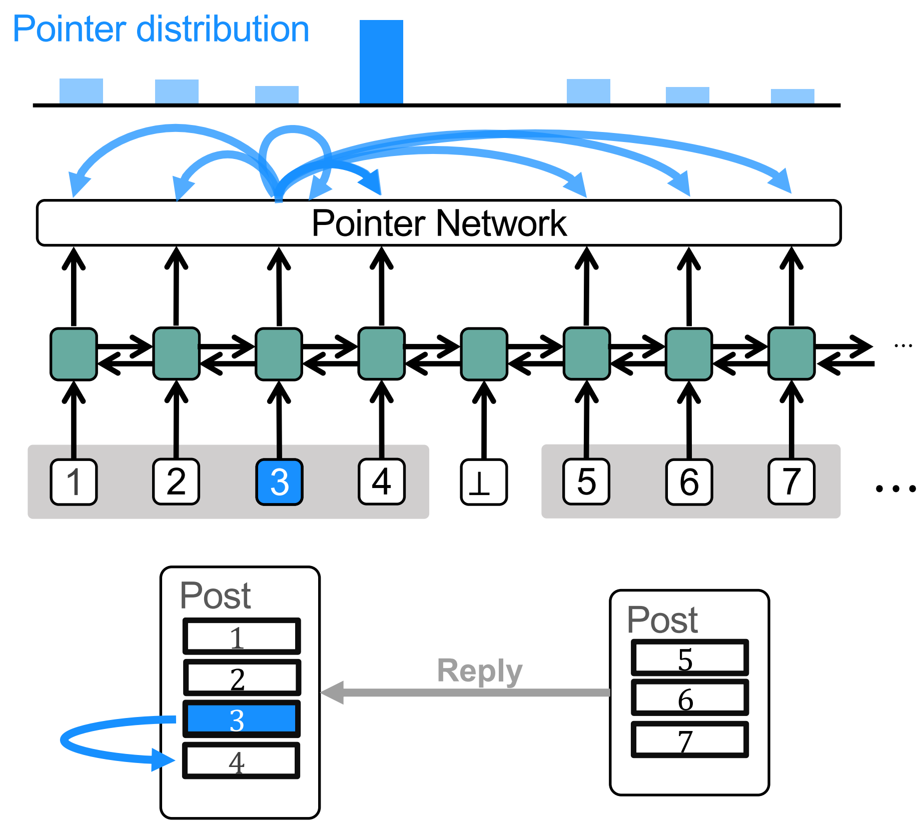


PCPA is composed of:

1. Input module
2. Encoding module
3. **Output modules**

e.g.

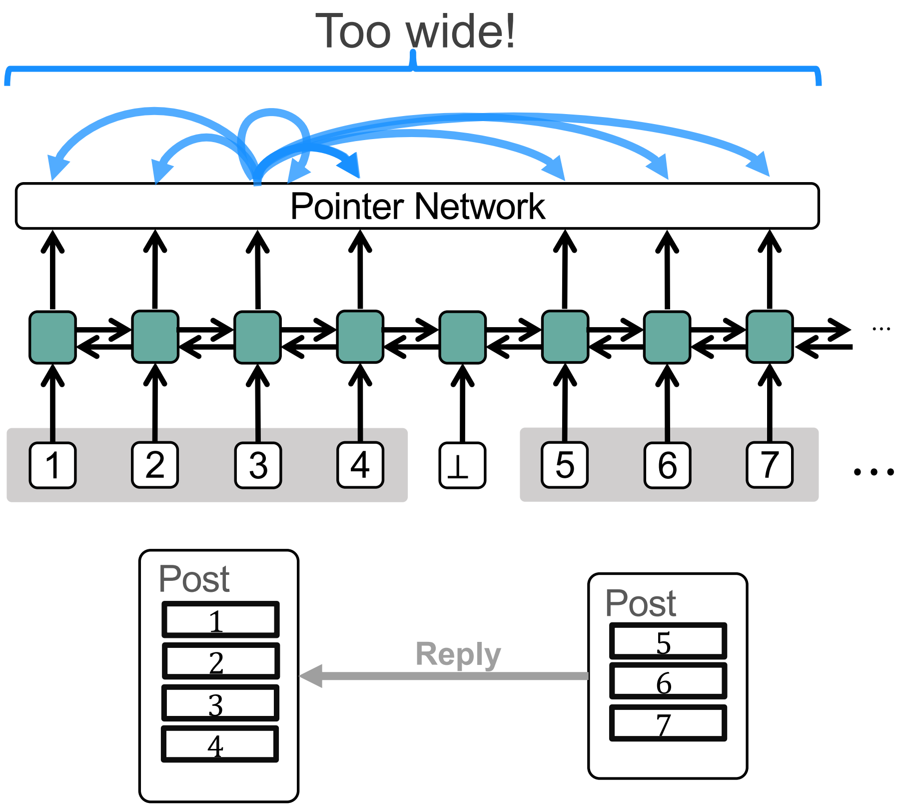
In this case, the IPR target is “**4**.” with the max value of the pointer distribution.



PCPA is composed of:

1. Input module
2. Encoding module
3. **Output modules**

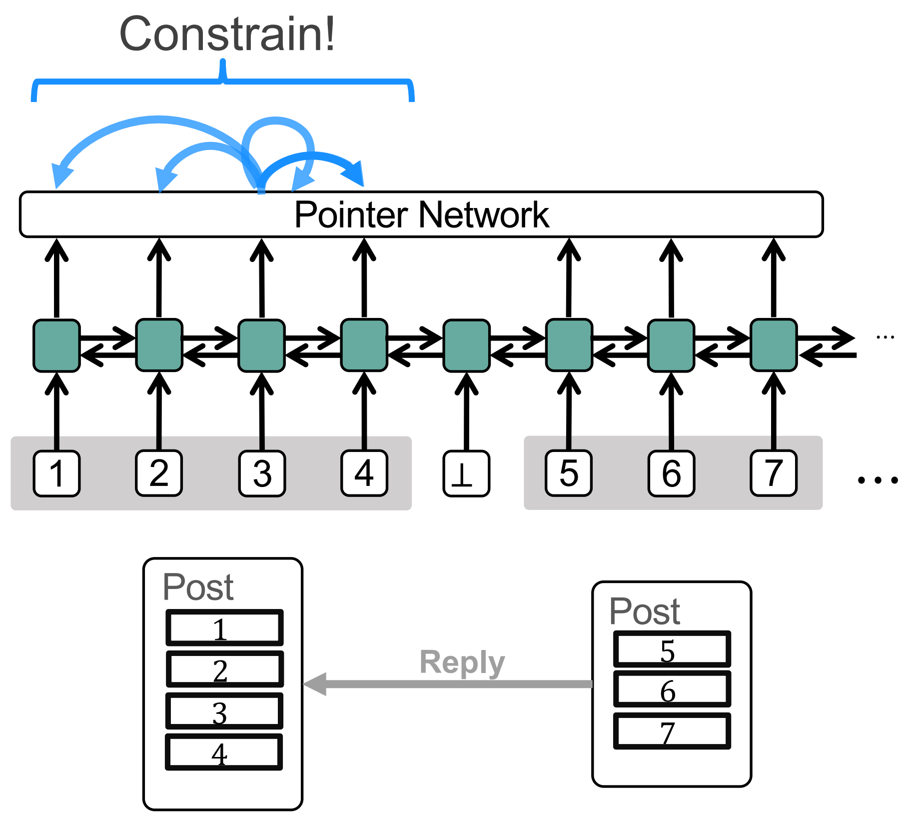
There is a problem;
we noticed that the computation space of an ordinal Pointer Network is too wide for our scheme.



PCPA is composed of:

1. Input module
2. Encoding module
3. **Output modules**

Therefore, PCPA constrains computation space. More specifically, **we don't need to scan out of post distributions** in IPR **because IPR is an inner-post relation**.



$$L_i^{ipr} = \sum_{j=1}^{N_i} \sum_{k=1}^{N_{i,j}} \log p(y_k^{ipr} \mid P_j^{(i)})$$

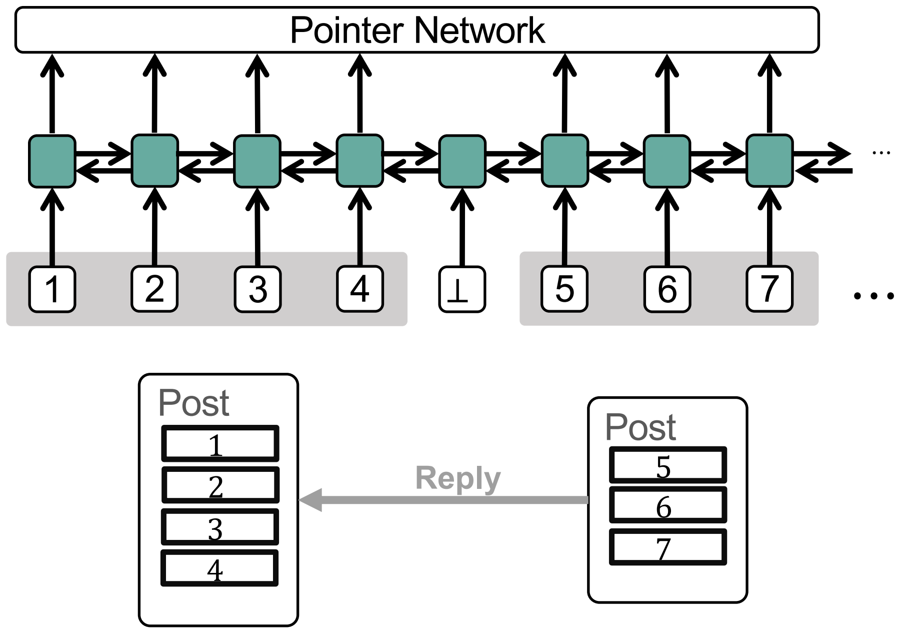
Objective

PCPA is composed of:

1. Input module
2. Encoding module
3. **Output modules**

3 IPI Classifier

Finally, we explain the inter-post interaction (IPI) layer.

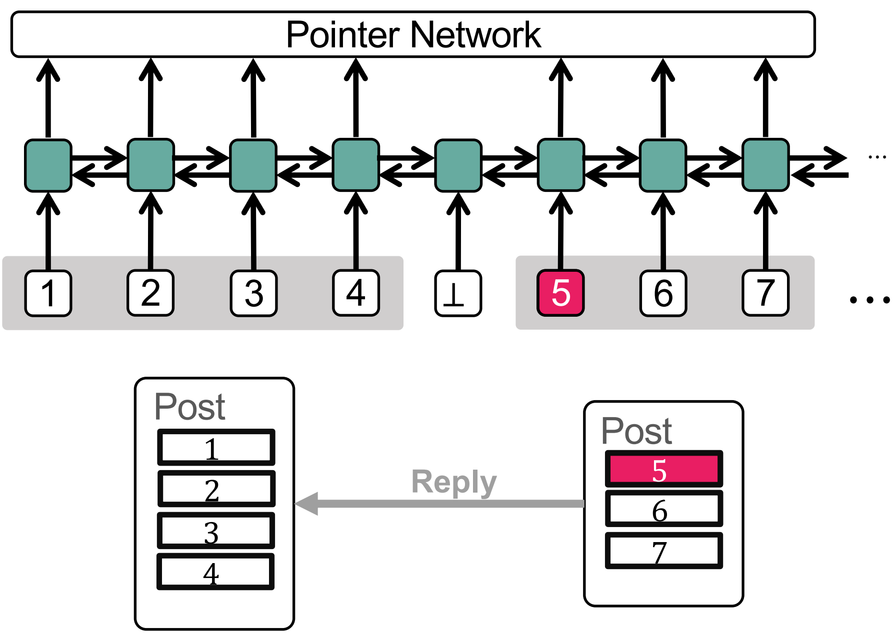


PCPA is composed of:

1. Input module
2. Encoding module
3. **Output modules**

e.g.

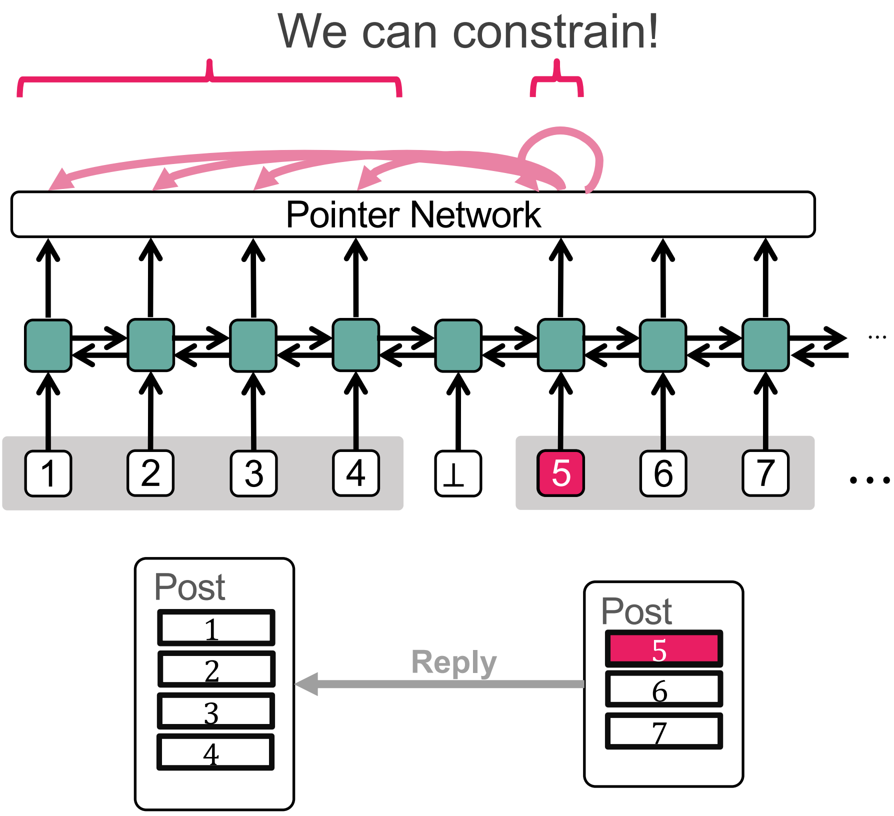
For the IPI classifier, we employ a pointer network similar to the IPR.
For example, let's search IPI target from sentence "5."



PCPA is composed of:

1. Input module
2. Encoding module
3. **Output modules**

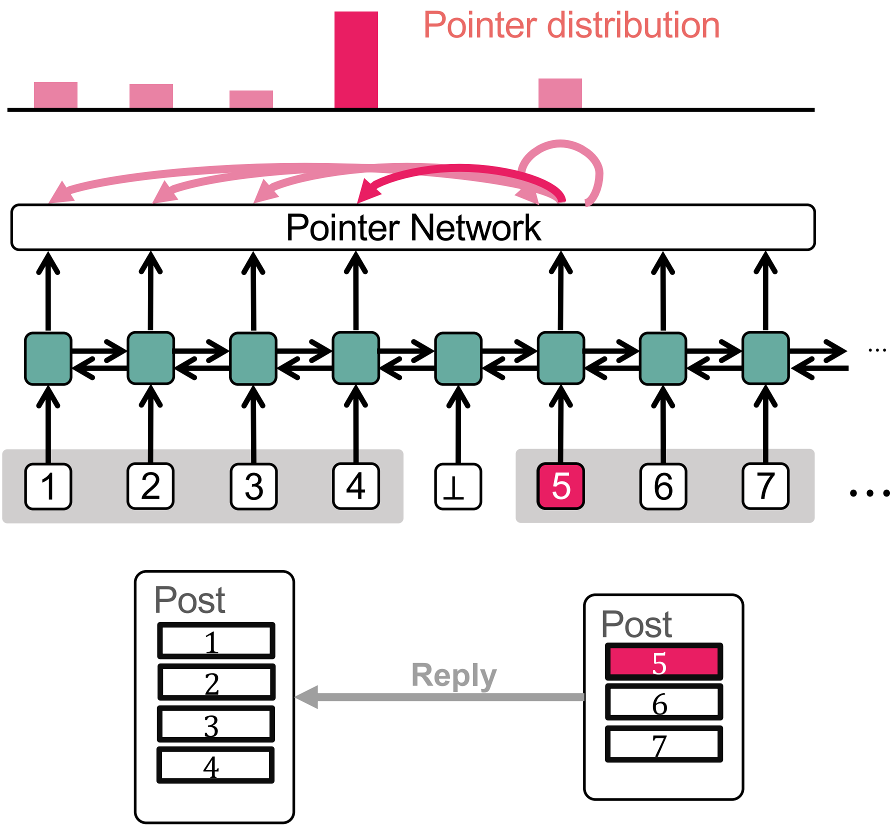
In the IPI, PCPA can also constrain computation space, and **we don't need to scan no relevant sentences** like "6,7" **because IPI is a post-to-post relation**.



PCPA is composed of:

1. Input module
2. Encoding module
3. **Output modules**

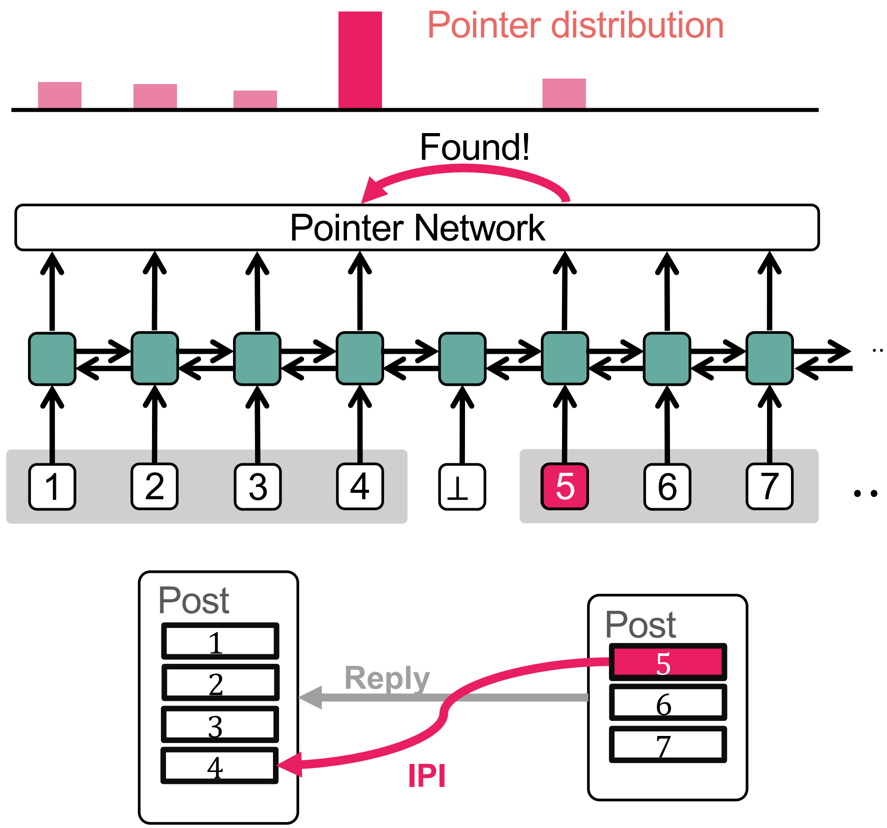
In the IPI, PCPA can also constrain computation space, and **we don't need to scan no relevant sentences** like "6,7" **because IPI is a post-to-post relation**.



PCPA is composed of:

1. Input module
2. Encoding module
3. **Output modules**

In the IPI, PCPA can also constrain computation space, and **we don't need to scan no relevant sentences** like "6,7" **because IPI is a post-to-post relation**.



$$p(y_k^{ipi} \mid P_{j_1}^{(i)}, P_{j_2}^{(i)}) = \mathrm{softmax}([q^{(i,j_1,k)}; q_k^{(i,j_2,k)}])$$

$$L_i^{ipi} = \sum_{(j_1,j_2)\in R^{(i)}} \sum_{k=1}^{N_{i,j_2}} \log p(y_k^{ipi} \mid P_{j_1}^{(i)}, P_{j_2}^{(i)})$$

Objective

PCPA is composed of:

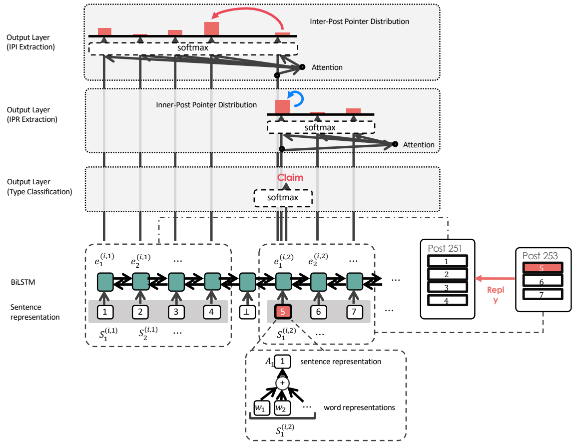
1. Input module
2. Encoding module
3. **Output modules**

$$Loss = \frac{1}{N}\sum_{i}(-\alpha L_i^{ipr} - \beta L_i^{ipi} - (1-\alpha-\beta)L_i^{type})$$

Finally, we arrive at the **final objective function.**

Time complexity

- PCPA **reduces** its time complexity compared to the standard Pointer Networks.
  - Given;
    - The average # of posts in a thread (n_p);
    - The average # of sentences in a post (n_S),
  - PCPA's time complexity is $O\left(n_s^2 * n_p\right)$ while the standard Pointer Networks take $O\left(n_s^2 * n_p^2\right)$.
    - You may think $O\left(n_s^2 * n_p\right)$ is large enough, though, the number of sentences per post is not so large in real world.



EXPERIMENTS

Experimental setting

- We employ following state-of-the-art baselines;
  - [Potash 2017] **Pointer Networks (Seq2Seq)**
    - An ordinal Pointer Networks (w/o constraints.)
  - [Potash 2017] **Pointer Networks (no Seq2Seq)**
    - Non- sequence-to-sequence model.
  - **MTL-BiLSTM** similar to [Eger 2017]
    - BiLSTM-based multi-task learning model which doesn’t employ Pointer Networks.

- Our dataset is split into, *train:test = 8:2.*

Performance results

- We show **F1 scores** for each model.
  - We can find from the table that **PCPA significantly outperforms all baselines in terms of IPR and IPI** classifications.
  - This results indicate that **constraining computation space is effective**.

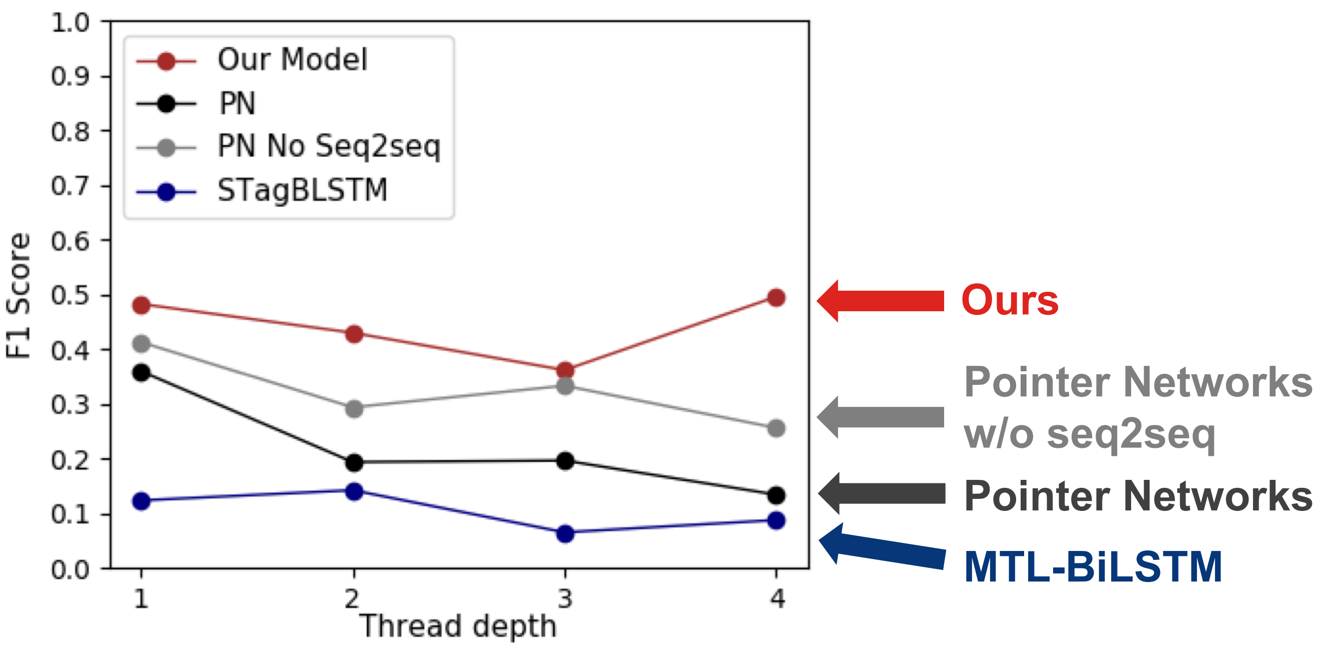
	Claim F1	Premise F1	NA F1	IPR F1	IPI F1
PCPA (ours)	58.1	**71.5**	**58.8**	***44.3**	***26.9**
Pointer Network (Seq2Seq)	58.3	70.8	48.6	27.2	19.4
Pointer Network (no Seq2Seq)	**60.1**	71.3	53.1	35.0	20.8
MTL-BiLSTM	54.2	65.6	56.9	14.9	12.6

For each model, we show the best score, and * indicates significant. at $p < 0.01$, two-sided Wilcoxon signed rank test.

IPR performance according to the **thread depth**

- We in turn observe performances of **inner-post relation (IPR)**, according to the thread depth.
- **In deeper threads**, ordinal Pointer Networks (PNs) can't keep their performances.
  - In contrast, our PCPA (red) **can keep the performance even for deeper threads**.

F1 for IPR.
↑

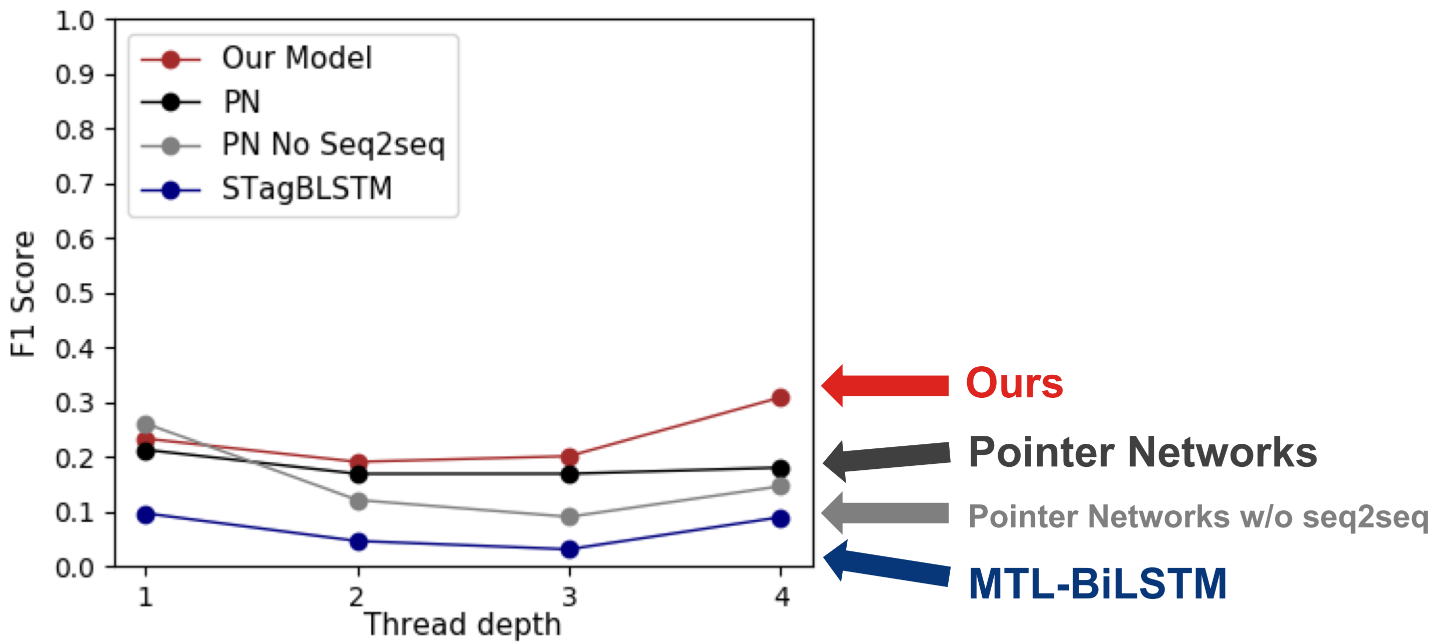


→ Thread depth.

IPI performance according to the **thread depth**

- For **inter-post interaction (IPI)**, our PCPA (red) **improves** the F1 score for deeper threads.

F1 for IPI. ↑



→ Thread depth.

CONCLUSION

Conclusion

1 We applied **Argument Mining for discussion threads.**

- Our scheme is based on [Stab 2017] and [Ghosh 2014].

2 We conducted **annotations for discussion threads.**

- Real online civic discussions are annotated.
- Inter-annotator agreements are evaluated.

3 We propose **P**arallel **C**onstrained **P**ointer **A**rchitecture

- The PCPA effectively constrains its computation space, and reduces time complexity.

4 **Experimental results demonstrate;**

- PCPA outperformed baselines significantly.
- Constraining computation space is effective for classifying the inner-post relation (IPR) and inter-post interaction (IPI).

ABOUT OUR DATA

Statistics of COLLAGREE data

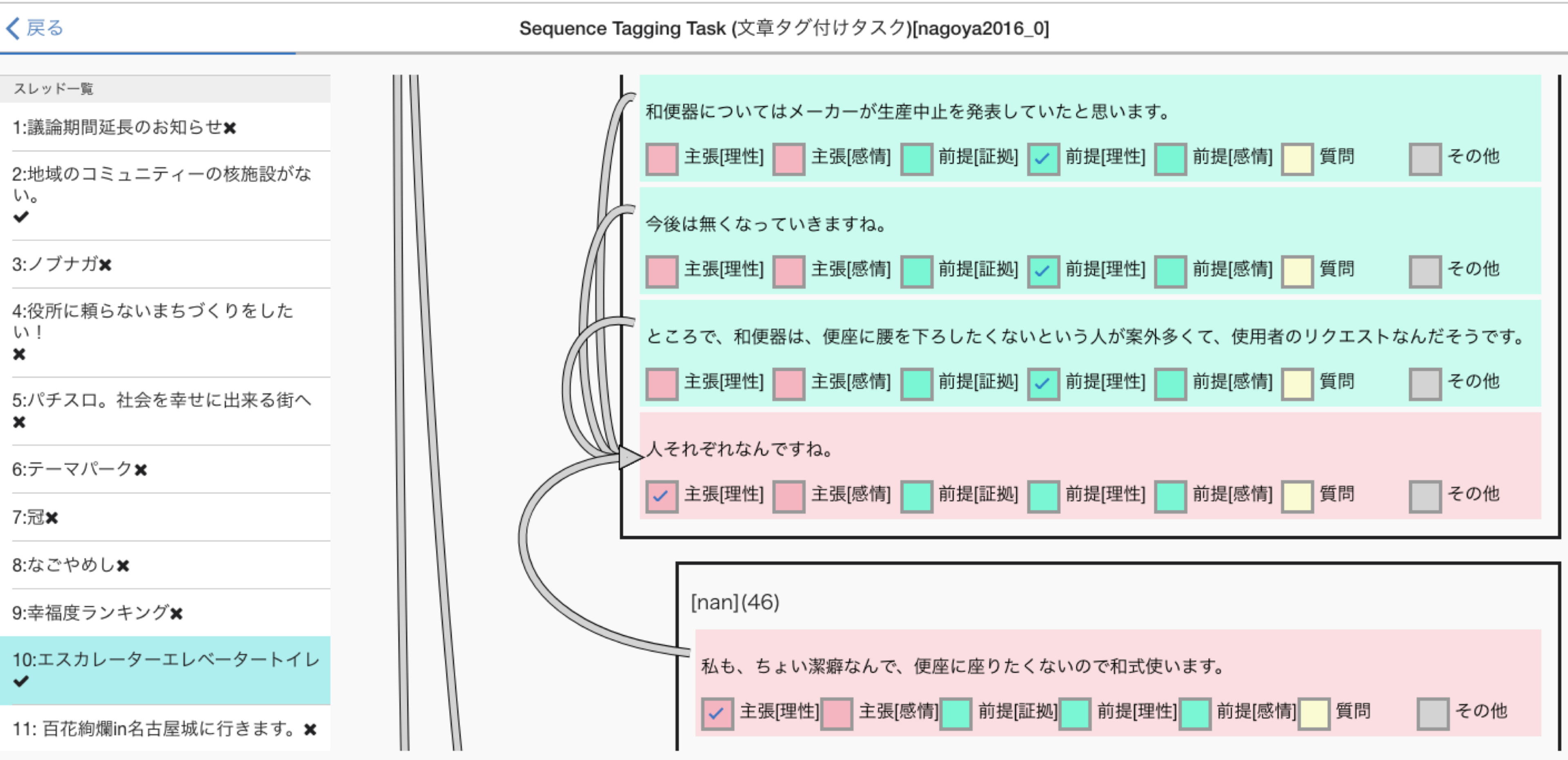
- **About COLLAGREE data**
  - Date: from 12.2016 to 1.2017
  - **204** citizens joined
  - **399** threads
  - **1327** threads
  - **5559** sentences

- **Average statistics**:
  - # of posts per thread: **3.33** （SD 3.29）
  - The depth of a thread: **1.09** （SD 1.19）
  - # of sentences per post: **4.19** （SD 3.33）
  - # of words per sentence: **21.63** （SD 19.92）

Statistics of COLLAGREE data

- **Annotation design**
  - Independent three annotators annotate each sentence.
    - **Annotation phase1** includes classifying each sentence into component types i.e., claim, premise and non-argumentative, and extracting support/attack relationships between them.
    - **Annotation phase2** includes extracting target/callout relationships between post-to-post interaction.
  - We evaluate kappa agreement using Fleiss' kappa.

Annotation Tool

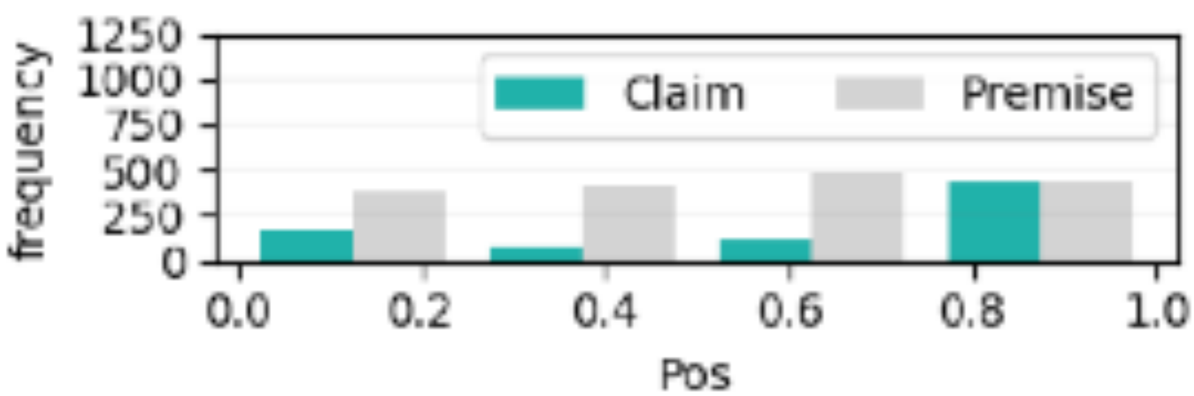


Positions of claim and premise in a post

- We examined position of argument components.

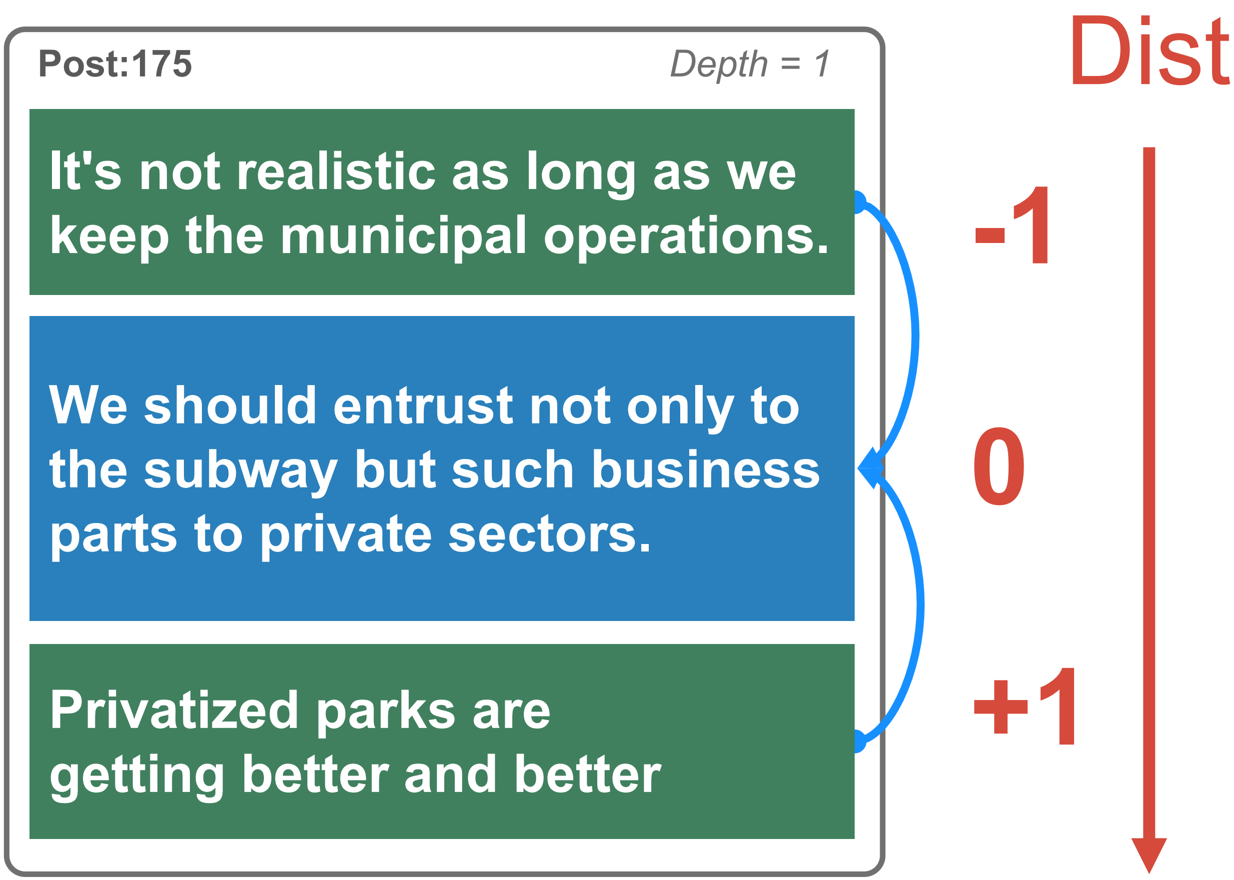
Positions of claim and premise in a post

- This figure below shows a histogram of position of premises and claims in posts with more than two sentences.
  - **Claims are tend to appear in the last of the post because citizens are likely to conclude their idea in the last.**



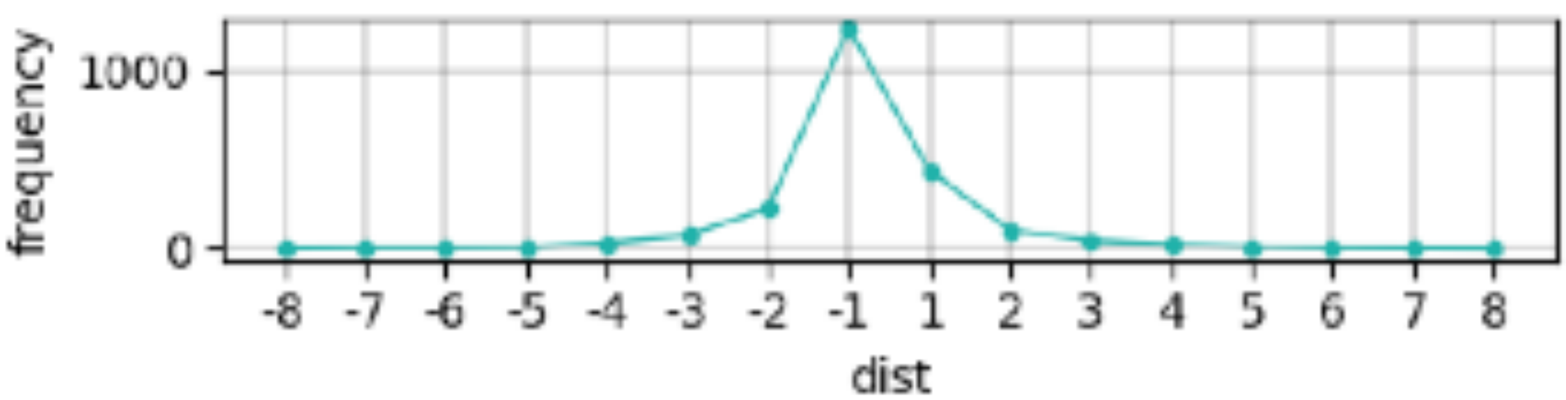
Premises’ distance from a claim

- We examined the distance of premises from a claim.



Premises' distance from a claim

- This figure below shows a histogram of premise Dist.
  - It shows that **premises are likely to appear immediately prior to a claim**.
  - In fact, the result exhibits the same property on the essay corpus [Eger 2017].



Distinct feature: IDF

- We investigate a histogram of the average Inversed document frequency (IDF) value per argument component (claim and premise) with more than 5 words.
  - The significance of averages shows at p < 0:0001.