The Seventeenth NTCIR conference (NTCIR-17)

KANOLAB

KIS's Stance Classification Model at the NTCIR-17 QA Lab-PoliInfo-4

Akira Nakada

Yoshinobu Kano

Shizuoka University

1st place in Stance Classification-2 task

- **QA Lab-Poliinfo4 Stance Classification-2 (SC2)**
  - **Formal run results**
    - **We achieved the 1st place 🥇🎉**

Rank	Team name	Accuracy（Leader Board）
#1	**KIS**	**0.9741**
#2	IKM23	0.9656
#3	ISLab	0.9326
#4	AKBL	0.9308

* Best scoring value for each team

Participation task overview

QA Lab-Poliinfo4 Stance Classification - 2 (SC2) [1]

Task overview

- estimate a politicians' **stance** from her/his utterances.

- **politician**
- **Topic** etc.

- Predict **stance**
  Agreement OR **Disagreement**

- **Utterance**

それでは、日本共産党議員団を代表いたしまして、議案第６号に **[STANCE]** する討論をします。この議案は、市会議員のボーナスを年間３．２５カ月から３．３０カ月と引き上げるものとなっています。
…
とても認めることができませんので、私ども日本共産党議員団は、議案第６号に **[STANCE]** をいたします。

* An utterance by a politician whose explicit tokens ("賛成" or "反対") are replaced with [STANCE]

Our Stance Classification Model

- **Our main methods**
  1. **Dealing with long utterances**
     - Using **head-tail method** [2]
  2. **Adaptation to the political domain**
     - Using **Incremental pretraining**
       - **Domain-adaptive pretraining（DAPT）[3]**
       - **Task-adaptive pretraining（TAPT）[3]**

Dealing with long utterances

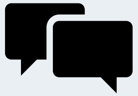
Distributed dataset characteristics

- ① More than **25 percent** exceed **512 tokens**.
  - * When using the LUKE-large tokenizer
- ② Important information is clustered at **the start or the end of an utterance**.

Head-tail method [2]

- **Uses only the specified number of tokens from the start and the end when truncating sentences**

● **Utterance**



それでは、日本共産党議員団を代表いたしまして、議案第６号に **[STANCE]** する討論をします。
この議案は、市会議員のボーナスを年間３．２５カ月から３．３０カ月と引き上げるものと
なっています。
....
とても認めることができませんので、私ども日本共産党議員団は、議案第６号に **[STANCE]** を
いたします。

Adaptation to the political domain

- **Existing models pretrained by corpora of the general domain (Wikipedia, etc.)**
  1. **Not much spoken language text included.**
  2. **Not much political domain text included.**

- here is currently no open language model specific to Japanese politics

Construction of Japanese politics-specific model

Adaptation to the political domain

- **Construction of Japanese politics-specific model**
  1. **Pretraining from scratch**
  2. **Incremental pretraining** ◀ **Our team choice**

- **Incremental pretraining**
  - **Incremental pretraining on original pretrained models**
    - **Using Diet minutes datasets**

Adaptation to the political domain

- **Diet minutes**
  - **Our team collected**
    - **House of Representatives from 2000 to 2022**
    - **House of Councillors from 2003 to 2022**

- Composed of **"spoken language"** not often included in general corpus
- Not enough data to pretrain from scratch.

Construction of political domain dataset

Adaptation to the political domain

- **Construction of political domain's dataset**
  - **Minute utterances are divided into blocks by 512 tokens**
  - **Two types of dataset:**

1. **Dataset of the entire minutes**
   - **694,907 blocks**
2. **Dataset of the discussion part**
   - **extracting utterances that contain a Japanese word "discussion" ("討論").**
   - **13,204 blocks**

Adaptation to the political domain

- Incremental pretraining
  - Domain-adaptive pretraining（DAPT） [3]
    - Adapting to the domain to which the task belongs
  - Task-adaptive pretraining（TAPT） [3]
    - Adapting to the task’s unlabeled data of training dataset

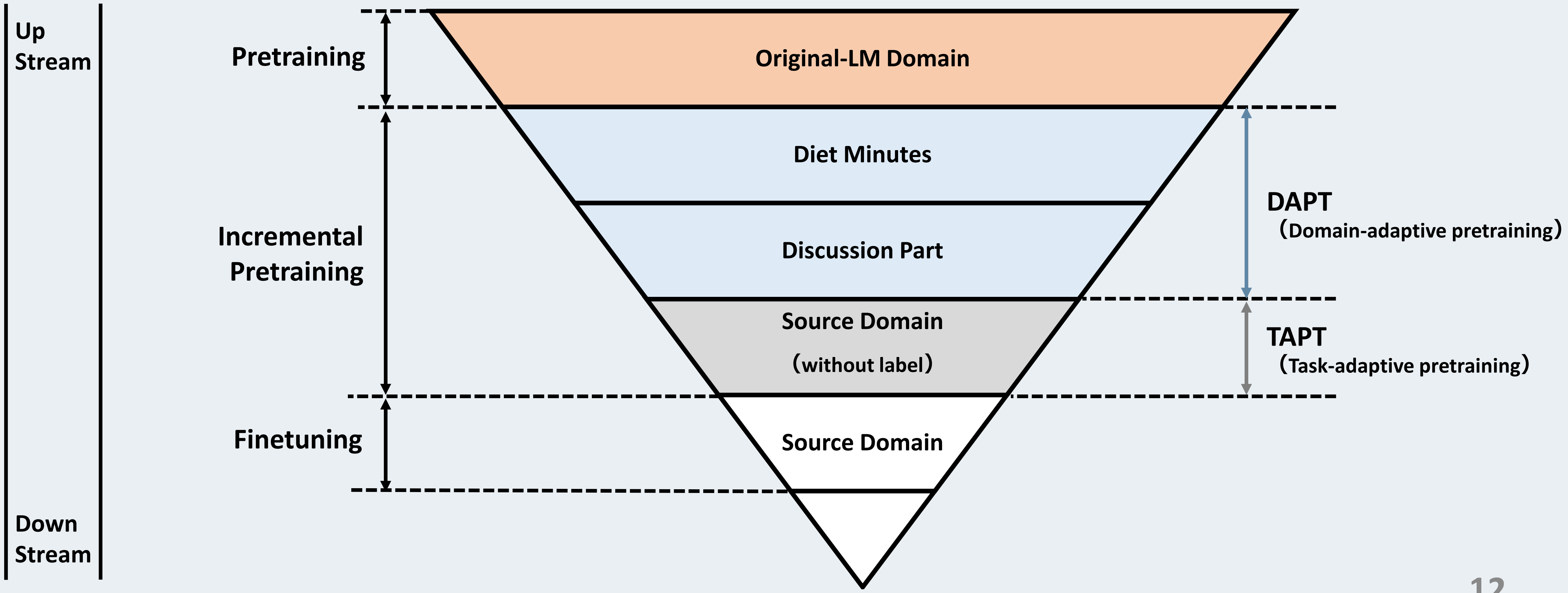
- Pretraining method used
  - Masked Language Model （MLM） [4]

Adaptation to the political domain

- **Adaptation to the political and task domain**
  - **Politics-specific model**
    - **Domain-adaptive pretraining（DAPT）**
      1. **DAPT1 (Minutes-specific model)**： MLM using the entire **Diet minutes** dataset
      2. **DAPT2 (Discussion-specific model)**： MLM using **the discussion part of the Diet minutes** dataset
  - **Task-specific model**
    - **Task-adaptive pretraining（TAPT）**
      1. TAPT： MLM using only the utterance portion of the SC2 task dataset without utilizing any labels

Adaptation to the political domain

- **Adaptation to the political domain**



Adaptation to the political domain

- **Fine tuning**
  - **K-fold cross validation（k=5）**
- **The final prediction**
  - **Majority vote**
    - **Prediction is determined by an ensemble of the predictions of all of the k fold models.**

- **Improved reliability** of model performance evaluation
- **Reduction of overtraining**

Experiments and Results

Selection of pretrained model

- **4 types of pretrained models: BERT-base and BERT-large, LUKE-base, LUKE-large**

model	5-fold Cross Validation			Leader board(Acc)
	Acc(max)	Acc(min)	Acc(avg)	
BERT-base	0.9274	0.9127	0.9191	0.9304
BERT-large	0.9268	0.9004	0.9139	0.9254
LUKE-base	0.9390	0.9285	0.9341	0.9469
LUKE-large	**0.9549**	**0.9349**	**0.9453**	**0.9563**

▶ **LUKE-large** showed a best performance in both validation and test data.

Experiments and Results

- **Dealing with long sentences**
  - **5 type of sentence truncation method**

base model	Head + tail ratio	5-fold Cross Validation			Leader board(Acc)
		Acc(max)	Acc(min)	Acc(avg)	
LUKE-large	512 / 0	0.9549	0.9349	0.9453	0.9563
	384 / 128	**0.9654**	0.9443	**0.9568**	0.9621
	256 / 256	0.9555	**0.9496**	0.9531	**0.9652**
	128 / 384	0.9596	0.9420	0.9502	**0.9652**
	0 / 512	0.9653	0.9436	0.9531	0.9612

▶ The combination of **384 tokens** and **128 tokens** showed a stable performance.

Experiments and Results

Effectiveness of Domain-Adaptive Pretraining（DAPT）

base model	model	5-fold Cross Validation			Leader board(Acc)
		Acc(max)	Acc(min)	Acc(avg)	
LUKE-large	without DAPT	0.9654	0.9443	0.9568	0.9621
	DAPT1	0.9695	**0.9566**	**0.9630**	0.9705
	DAPT2	0.9590	0.9490	0.9551	0.9610
	DAPT1 + DAPT2	**0.9672**	0.9537	0.9625	**0.9728**

- **DAPT1** showed better performance than the base model in both validation and test data.
- **DAPT1 + DAPT2** showed a best performance in test data.

Experiments and Results

Effectiveness of Task-Adaptive Pretraining（TAPT）

base model	model	5-fold Cross Validation			Leader board(Acc)
		Acc(max)	Acc(min)	Acc(avg)	
LUKE-large	without DAPT	0.9654	0.9443	0.9568	0.9621
	TAPT	**0.9883**	**0.9830**	**0.9852**	0.9629
	DAPT1 + TAPT	0.9672	0.9596	0.9633	0.9696
	DAPT1 + DAPT2 + TAPT	0.9736	0.9602	0.9652	**0.9741**

- **TAPT** showed slightly better in test data.
- **DAPT1 + DAPT2 + TAPT** showed a best performance in test data.

Error Analysis

- Error analysis
  - Analysis of incorrectly answered utterances
    - Statistical Analysis
    - Qualitative Analysis

- Our best model was correct for **about 97%** of the test data

What utterances were misclassified?

Error Analysis

- **Statistical Analysis**
  - **Number of samples of incorrect answers of the test data**
    - **58 / 2240 samples (2.6%)**
  - **Token Length Statistics**

	Incorrect data	Correct data
Mean	494.19	455.90
Median	394.5	361.0
Percentage(> 512)	32.76(%)	27.31(%)

* When using the LUKE-large tokenizer

▶ **Possibility that important information was lost during truncation.**

Error Analysis

- **Qualitative Analysis**
  - **Characteristics of incorrect answers (subjective assessment)**
    - **Cases of partial agree (or disagree) in utterance**
    - **Cases in which the inverse conjunction is used**
      - e.g., **However**, **But**

- Our model only saw the **start** and the **end** of the sentence.

Therefore, could not recognize the switch in stance.

Error Analysis

Qualitative Analysis

- **A prominent example.** ***English translation version follows in the next slide.**

Stance: **Disagreement**

● Utterance

それでは、まず第一に、13万人市民の命を市長初め市の職員の方が守っていただき、**大変ありがとうございます。**
（中略）
しかしながら、地方自治体の両輪の輪としてチェック機関としてちょっと厳しい意見を述べますが、それは応援メッセージという形で聞いてください。
[STANCE]の立場で討論いたします。
（中略）

id = 112275-00373

Error Analysis

Qualitative Analysis

A prominent example

Stance: **Disagreement**

● **Utterance**

First, **I would like to thank the mayor and the city officials** for protecting the lives of 130,000 citizens.
(omitted)
However, as a checking body in the two wheels of local government, **I would like to express a little harsh opinion**, but please hear it in the form of a message of support. I will debate from the standpoint of **[STANCE]**.
(omitted)

id = 112275-00373

Stance: Agreement -> Disagreement (The stance is switched in the middle of the utterance.)

Conclusion

- **Dealing with long utterances by head-tail method**
  - **Good performance in situations where a politician's utterances is used.**
- **Japanese politics-specific model that we constructed**
  - **Good performance in SC2 task**

Conclusion

- **Future work**

1. **Improve model performance based on error analysis**
   - Dealing with long utterances
2. **Adaptation of the politics-specific model to other tasks**
   - Other NTCIR17-QA-Lab Poliinfo4 Tasks
   - Analysis of political-related social networking posts

REFERENCES

- [1]: Yasuhiro Ogawa, Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Kei- ichi Takamaru, Kazuma Kadowaki, Tomoyoshi Akiba, Minoru Sasaki, Akio Kobayashi, Masaharu Yoshioka, Tatsunori Mori, Kenji Araki, Satoshi Sekine, and Teruko Mitamura. Overview of the ntcir-17 qa lab-poliinfo-4 task. Proceedings of The 17th NTCIR Conference, 12 2023.

- [2]: Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to Fine-Tune BERT for Text Classification?. In Chinese Computational Linguistics, Maosong Sun, Xuanjing Huang, Heng Ji, Zhiyuan Liu, and Yang Liu (Eds.). Springer International Publishing, Cham, 194–206.

- [3]: Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740

- [4]: Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423