# HCLT2021

# An Empirical Study of Topic Classification for Korean Newspaper Headlines

**Jeiyoon Park,** Mingyu Kim, Yerim Oh, Sangwon Lee, Jiung Min, Youngdae Oh
Presentation: Jeiyoon Park

# Outline

1. Contribution
2. Experiments
3. Conclusion

# Outline

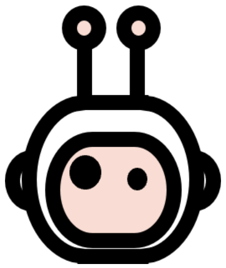
1. Contribution
2. Experiments
3. Conclusion

# Detour: Why NLU?

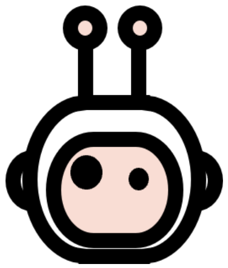
## 1. Ambiguity

- There are several levels at which ambiguity may occur in natural language: *Syntactical, lexical, referential, semantic, and pragmatic level*

- e.g.) “I hit the robot with the hammer”



I hit the robot with the hammer

A robot must obey the orders given it

# Detour: Why NLU?


## 1. Ambiguity

- As humans we are adept at coping with these things, to the extent that we can usually understand each other if we speak the same language, even if words are missed out or misused.

- We usually have enough knowledge in common to disambiguate the words and interpret them correctly in context.

- We can also cope quickly with new words. This is borne out by the speed with which slang and street words can be incorporated into everyday usage.

# Detour: Why NLU?

## 2. Applications

- Topic classification
- Named entity recognition
- Question answering
- Machine translation
- Text Summarization
- Machine reading comprehension

→ Large-scale language model

# Contribution

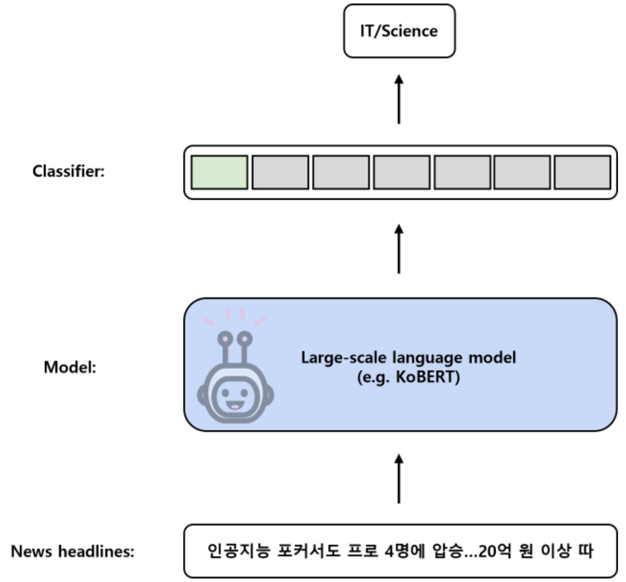
## An empirical study of Topic classification

- A good and human-like natural language understanding system should infer what the text means, not only recognize the shape of a word or sentence in a text.

- We compared the performance of various Korean large-scale models that have not been previously tested for comparison, and empirically analyzed the causes of the results.

- Open benchmark:
*KLUE-TC*

- Off-the-shelf baselines:
*KoBERT, KoBART, KoELECTRA, and KcELECTRA*



# Outline

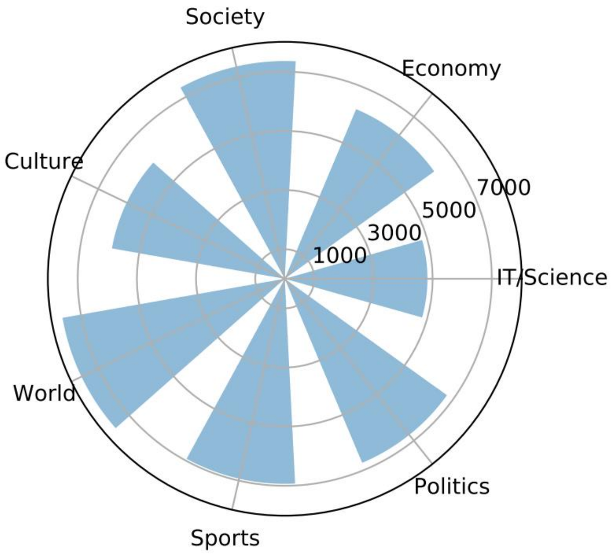
1. Contribution
2. Experiments
3. Conclusion

# Experiments

## 1. Benchmark: KLUE-TC

- It collected by Yonhap News Agency from Jan 2016 to Dec 2020.
- Each headline is classified into seven classes: IT/Science, Economy, Society, Culture, World, Sports, and Politics.

| Topic | Label | Headline Examples | #Training | #Test |
|---|---|---|---|---|
| IT/Science | 0 | 내년 첫 5G폰 평균가 80만원 육박...2023년 60만원대↓ | 4,824 | |
| Economy | 1 | 코스피 기관·개인 매수에 반등...코스닥 1%대 상승종합 | 6,222 | |
| Society | 2 | 뜨거운 감자 포털 규제...필요성·시행안 등 논쟁 가열종합 | 7,362 | |
| Culture | 3 | 웹툰 나이트에 참가한 작가들 | 5,933 | |
| World | 4 | 총기폭력 더 방치 못해...美 초강력 총기규제안 도입종합 | 7,629 | |
| Sports | 5 | 아시안피스컵 남북 男배구 열전 펼쳐...북한팀 32 승종합 | 6,933 | |
| Politics | 6 | 국회사무처 美매사추세츠大·뉴욕시립대 퀸즈칼리지와 MOU | 6,751 | |
| **Total** | | | **45,654** | **9,131** |



# Experiments

## 2. Baselines: KoBERT, KoBART, KoELECTRA, KcELECTRA

1) **KoBERT**

- Korean wiki (# 5M sentences), # parameters: 92M

2) **KoBART**

- Korean wiki (# 5M sentences) + other corpus (0.27B), # parameters: 124M

3) **KoELECTRA**

- Korean news, wiki, and namuwiki (14G + 20G), # parameters: *110M (note: not official, ELECTRA-base)*

4) **KcELECTRA**

- Naver news comments (17.3G), # parameters: *110M (note: not official, ELECTRA-base)*

# Experiments

## 3. Results: KLUE-TC and NSMC task

- NSMC (left) and KLUE-TC (right)

| Method | Accuracy |
| --- | --- |
| KoBART | 90.24 |
| KoELECTRA | 90.63 |
| KcELECTRA | **91.54** |
| KoBERT | 89.59 |

| Method | Accuracy |
| --- | --- |
| KoBART | 85.62 |
| KoELECTRA | 84 |
| KcELECTRA | 84.57 |
| KoBERT | **86.7** |

- **KoELECTRA**: Korean news, wiki, and namuwiki (14G + 20G -> about 0.18B sentences), # parameters: 110M
- **KoBERT**: Korean wiki (# 5M sentences), # parameters: **92M**

# Outline

1. Contribution
2. Experiments
3. Conclusion

# Conclusion

## 1. Analysis

(1) KoBERT achieves highest score among baselines (86.7), despite guaranteed pretraining method and diverse and bigger corpus

(2) This result demonstrates that we should employ training more cautiously, for instance, the noise generation method used for pretraining the Korean language model and constructing the corpus

# Conclusion

## 2. Future work

(1) The other KLUE benchmarks
(2) KLUE-BERT, KLUE-RoBERTa

| **Model** | **SQuAD 1.1/2.0** | **MNLI-m** | **SST-2** | **RACE** |
|---|---|---|---|---|
| *Our reimplementation (with NSP loss):* | | | | |
| SEGMENT-PAIR | 90.4/78.7 | 84.0 | 92.9 | 64.2 |
| SENTENCE-PAIR | 88.7/76.2 | 82.9 | 92.1 | 63.0 |
| *Our reimplementation (without NSP loss):* | | | | |
| FULL-SENTENCES | 90.4/79.1 | 84.7 | 92.5 | 64.8 |
| DOC-SENTENCES | 90.6/79.7 | 84.7 | 92.7 | 65.6 |
| $\mathrm{BERT}_{\mathrm{BASE}}$ | 88.5/76.3 | 84.3 | 92.8 | 64.3 |
| $\mathrm{XLNet}_{\mathrm{BASE}}$ (K = 7) | –/81.3 | 85.8 | 92.7 | 66.1 |
| $\mathrm{XLNet}_{\mathrm{BASE}}$ (K = 6) | –/81.0 | 85.6 | 93.4 | 66.7 |

Thank you

https://jeiyoon.github.io/