KLUE : Korean Language Understanding Evaluation

Sungjoon Park et al. (+30 people)

Upstage, KAIST, etc. (+10 organizers)

NeurIPS, 2021

Nayeon Kim

lilian1208@korea.ac.kr
Korea University

Contents

Problem Statement

How can we evaluate

Korean language understanding ability of language models?

Background

2018

Google
BERT

2020

GPT3

Language model(LM)

2018

GLUE

2019

SuperGLUE

Natural Language Understanding(NLU) Benchmark Dataset

[1] Wang et al., GLUE : A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, ICLR, 2019
[2] Wang et al., SuperGLUE : A Stickier Benchmark for General-Purpose Language Understanding Systems, NeurIPS, 2019

Background

- **GLUE Leaderboard**

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MI
1	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	
2	Microsoft Alexander v-team	Turing NLR v5		91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	
3	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	
4	DIRL Team	DeBERTa + CLEVER		91.0	74.5	97.5	93.3/91.0	93.4/93.1	76.4/90.9	92.1	
5	AliceMind & DIRL	StructBERT + CLEVER		91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	
6	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	
21	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	

Language models outperform human performance.

Background

- GLUE Leaderboard

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	M
1	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	
[illegible]	Microsoft Alexander v-team	Turing N[illegible]		91.2	72.6	97.[illegible]	[illegible].8/91.7	93.[illegible]/93.3	[illegible].4/91.[illegible]	92.6	
3	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	
4	DIRL Team	DeBERTa + CLEVER		91.0	74.5	97.5	93.3/91.0	93.4/93.1	76.4/90.9	92.1	
5	AliceMind & DIRL	StructBERT + CLEVER		91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	
6	DeBERTa Team - Microsoft	[illegible]		[illegible]0.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	
21	GLUE Human Baselines	[illegible]		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	

We need more difficult language understanding tasks!

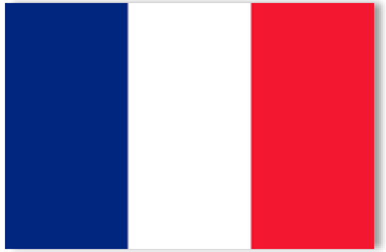
SuperGLUE

Language models outperform human performance.

Background

- **Language-specific benchmark**

Chinese GLUE



French GLUE

IndoNLU

Russian SuperGLUE

[1] Liang et al., CLUE : A Chinese language understanding evaluation benchmark, ACL, 2020
[2] Le et al., FlauBERT : Unsupervised language model pre-training for French, LREC, 2020
[3] Wilie et al., IndoNLU : Benchmark and resources for evaluating Indonesian natural language understanding, ACL, 2020
[4] Shavrina et al., RussianSuperGLUE : A Russian language understanding evaluation benchmark, EMNLP, 2020

Background

- **Pretrained Language Models for Korean**

KoBERT

KorBERT

KcBERT

HanBERT

KR-BERT

KoELECTRA

.
.
.

How can we compare them?

There is a lack of standard benchmark in Korean.

Background

- Pretrained Language Models for Korean

KR-BERT

KoELECTRA

·
·
·

How can we compare them?

There is a lack of standard

KLUE

Korean Language Understanding Evaluation

Contributions

- They build a new benchmark suite for evaluating NLU in Korean.
- They provide suitable evaluation metrics and fine-tuning recipes for pretrained language models for each task.
- They release large-scale pretrained language models for Korean.

KLUE

KLUE Benchmark

- **Design principles**

Diversity

Covering diverse tasks and corpora

Accessible

Accessible to everyone without any restriction

Accurate and unambiguous

For accurate performance evaluation

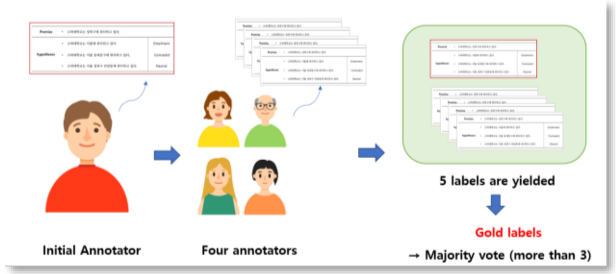
Ethical issues

Mitigating AI ethical issues

KLUE Benchmark

- **Building process**

Name	Type	Format	Eval. Metric	# Class	{|Train|, |Dev|, |Test|}	Source	Style
KLUE-TC (YNAT)	Topic Classification	Single Sentence Classification	Macro F1	7	45k, 9k, 9k	News (Headline)	Formal
KLUE-STS	Semantic Textual Similarity	Sentence Pair Regression	Pearson's r, F1	[0, 5] 2	11k, 0.5k, 1k	News, Review, Query	Colloquial, Formal
KLUE-NLI	Natural Language Inference	Sentence Pair Classification	Accuracy	3	25k, 3k, 3k	News, Wikipedia, Review	Colloquial, Formal
KLUE-NER	Named Entity Recognition	Sequence Tagging	Entity-level Macro F1 Character-level Macro F1	6, 12	21k, 5k, 5k	News, Review	Colloquial, Formal
KLUE-RE	Relation Extraction	Single Sentence Classification (+2 Entity Spans)	Micro F1 (without *no_relation*), AUPRC	30	32k, 8k, 8k	Wikipedia, News	Formal
KLUE-DP	Dependency Parsing	Sequence Tagging (+ POS Tags)	Unlabeled Attachment Score, Labeled Attachment Score	# Words, 38	10k, 2k, 2.5k	News, Review	Colloquial, Formal
KLUE-MRC	Machine Reading Comprehension	Span Prediction	Exact Match, ROUGE-W (LCCS-based F1)	2	12k, 8k, 9k	Wikipedia, News	Formal
KLUE-DST (WoS)	Dialogue State Tracking	Slot-Value Prediction	Joint Goal Accuracy Slot Micro F1	(45)	8k, 1k, 1k	Task Oriented Dialogue	Colloquial



1. Collecting base corpora

2. Identifying a set of benchmark tasks

3. Designing annotation protocol

4. Selecting qualified workers

5. Collecting annotations

6. Validating collected annotation

KLUE Benchmark

- Building process

1. Collecting base corpora

- News Headlines from Yonhap News Agency
- Wikipedia
- Wikinews
- Wikitree
- Policy News
- ParaKQC
- Airbnb Reviews
- NAVER Sentiment Movie Corpus
- The Korean Economics Daily News
- Acrofan News

KLUE Benchmark

- Building process

1. Collecting base corpora

Source Corpora

- News Headlines from Yonhap News Agency
- Wikipedia
- Wikinews
- Wikitree
- Policy News
- ParaKQC
- Airbnb Reviews
- NAVER Sentiment Movie Corpus
- The Korean Economics Daily News
- Acrofan News

✓ Derivative work
✓ Redistribution
✓ Commercial use

KLUE Benchmark

- **Building process**

1. Collecting base corpora

Name	Type	Format	Eval. Metric	# Class	{|Train|, |Dev|, |Test|}	Source	Style
KLUE-TC (YNAT)	Topic Classification	Single Sentence Classification	Macro F1	7	45k, 9k, 9k	News (Headline)	Formal
KLUE-STS	Semantic Textual Similarity	Sentence Pair Regression	Pearson's r, F1	[0, 5] 2	11k, 0.5k, 1k	News, Review, Query	Colloquial, Formal
KLUE-NLI	Natural Language Inference	Sentence Pair Classification	Accuracy	3	25k, 3k, 3k	News, Wikipedia, Review	Colloquial, Formal
KLUE-NER	Named Entity Recognition	Sequence Tagging	Entity-level Macro F1 Character-level Macro F1	6, 12	21k, 5k, 5k	News, Review	Colloquial, Formal
KLUE-RE	Relation Extraction	Single Sentence Classification (+2 Entity Spans)	Micro F1 (without *no_relation*), AUPRC	30	32k, 8k, 8k	Wikipedia, News	Formal
KLUE-DP	Dependency Parsing	Sequence Tagging (+ POS Tags)	Unlabeled Attachment Score, Labeled Attachment Score	# Words, 38	10k, 2k, 2.5k	News, Review	Colloquial, Formal
KLUE-MRC	Machine Reading Comprehension	Span Prediction	Exact Match, ROUGE-W (LCCS-based F1)	2	12k, 8k, 9k	Wikipedia, News	Formal
KLUE-DST (WoS)	Dialogue State Tracking	Slot-Value Prediction	Joint Goal Accuracy Slot Micro F1	(45)	8k, 1k, 1k	Task Oriented Dialogue	Colloquial

2. Identifying a set of benchmark tasks

3. Designing appropriate annotation protocol

1. Topic Classification(TC)
2. Semantic Textual Similarity (STS)
3. Natural Language Inference (NLI)
4. Named Entity Recognition (NER)
5. Relation Extraction (RE)
6. Dependency Parsing (DP)
7. Machine Reading Comprehension(MRC)
8. Dialogue State Tracking (DST)

KLUE Benchmark

- **Building process**

1. Collecting base corpora

Name	Type	Format	Eval. Metric	# Class	{|Train|, |Dev|, |Test|}	Source	Style
KLUE-TC (YNAT)	Topic Classification	Single Sentence Classification	Macro F1	7	45k, 9k, 9k	News (Headline)	Formal
KLUE-STS	Semantic Textual Similarity	Sentence Pair Regression	Pearson's r, F1	[0, 5] 2	11k, 0.5k, 1k	News, Review, Query	Colloquial, Formal
KLUE-NLI	Natural Language Inference	Sentence Pair Classification	Accuracy	3	25k, 3k, 3k	News, Wikipedia, Review	Colloquial, Formal
KLUE-NER	Named Entity Recognition	Sequence Tagging	Entity-level Macro F1 Character-level Macro F1	6, 12	21k, 5k, 5k	News, Review	Colloquial, Formal
KLUE-RE	Relation Extraction	Single Sentence Classification (+2 Entity Spans)	Micro F1 (without *no_relation*), AUPRC	30	32k, 8k, 8k	Wikipedia, News	Formal
KLUE-DP	Dependency Parsing	Sequence Tagging (+ POS Tags)	Unlabeled Attachment Score, Labeled Attachment Score	# Words, 38	10k, 2k, 2.5k	News, Review	Colloquial, Formal
KLUE-MRC	Machine Reading Comprehension	Span Prediction	Exact Match, ROUGE-W (LCCS-based F1)	2	12k, 8k, 9k	Wikipedia, News	Formal
KLUE-DST (WoS)	Dialogue State Tracking	Slot-Value Prediction	Joint Goal Accuracy Slot Micro F1	(45)	8k, 1k, 1k	Task Oriented Dialogue	Colloquial

2. Identifying a set of benchmark tasks

3. Designing appropriate annotation protocol

- Why include?
- Task Format
- Evaluation

→ Specific Example → Annotation protocol / process → Final dataset

1. Topic Classification(TC)

- **Why include TC?**
  - Inferring the topic of a text is a key capability that should be possessed by a language understanding system.
  - For Korean, no dataset has been proposed for the task.
- **Task Format**
  - Single Sentence Classification
- **Evaluation Metric**
  - Marco F1

1. Topic Classification(TC)

- **Source Corpora : Yonhap News Agency 2016.1 – 2020.12**

Politics

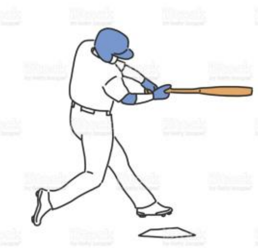
Economy

Society

Culture

World

IT/Science



Sports

7 sections → 70,000 headlines

10,000 articles from each section,
except for the sports & IT/science

9,000 11,000

1. Topic Classification(TC)

- **What is TC?**

연합뉴스TV

삼성전자 미성년 주주 35만명…1인당 41주 보유

입력2022.05.05. 오전 10:55 기사원문

추천 댓글

소위 '국민주'로 불리는 삼성전자 주식을 보유한 20대 미만 미성년 주주가 35만 명을 넘어선 것으로 나타났습니다.

한국예탁결제원에 따르면 지난해 말 기준 삼성전자의 20대 미만 주주는 35만8,257명으로 역대 최대 규모였습니다.

미성년 주주는 삼성전자 전체 주주 506만여명 중 7.07%를 차지했고, 전체 발행 주식의 0.25% 수준인 총 1,483만4,499주를 보유하고 있었습니다.

- **Politics**
- ✓ **Economy**
- **Society**
- **Culture**
- **World**
- **IT/Science**
- **Sports**

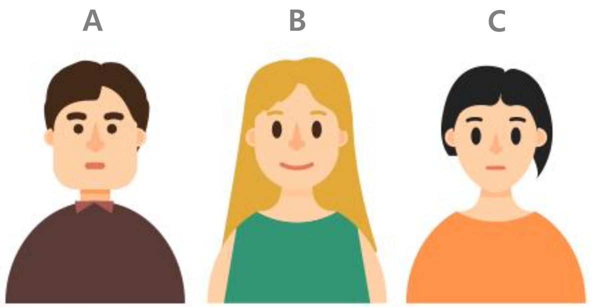
- **To train a classifier to predict the topic of a given text snippet.**

1. Topic Classification(TC)

- **Annotation Protocol**

<Key terms>

Politics : Blue House, Ministry, Defense / Society : Labor, Food and drugs, Human rights…



Annotators	First	Second	Third
A	Social	X	X
B	IT/Science	Social	X
C	IT/Science	X	X

1) They present *key terms* of each topic to annotators.

2) Three annotators label topics independently from each other.

3) They pick at most three topics among the seven categories.

1. Topic Classification(TC)

- **Annotation Protocol**

Annotators	Headlines	First	Second	Third	Final Label
A		Economy	X	X	
B	“삼성전자 미성년 주주 35만명… 1인당 41주 보유”	Social	Economy	X	Economy
C		Economy	X	X	

1. Topic Classification(TC)

- **Annotation Process**

✓ **13 selected workers labeled topics for all 70,000 headlines.**

✓ **Workers reported headlines :**

toxic contents, PII(Personally identifiable information**)s, unable-to-decides**

Filtered final Datasets : 63,892 (Train : 45,678 / Dev : 9,107 / Test : 9,107)

YNAT(Yonhap News Agency datasets for Topic Classification)

2. Semantic Textual Similarity (STS)

- **Why include STS?**

  - It is essential to other NLP tasks such as machine translation, summarization, and QA.

- **Task Format**

  - Sentence-Pair Regression

- **Evaluation Metric**

  - Persons's r
  - F1

2. Semantic Textual Similarity (STS)

▪ **What is STS?**

Sentence 1 **Sentence 2**	“지하철을 타도 30분 안에는 이동이 가능합니다! ” “지하철을 탄다고 해도, 30분이면 그곳에 도착할 수 있어요! ”	**Similarity : 4.0**
Sentence 1 **Sentence 2**	“위반행위 조사 등을 거부·방해·기피한 자는 500만원 이하 과태료 부과 대상이다.” “시민들 스스로 자발적인 예방 노력을 한 것은 아산 뿐만이 아니었다.”	**Similarity : 0.0**

- **To predict the semantic similarity of two input sentences as a real value from 0 to 5.**

2. Semantic Textual Similarity (STS)

- **Source Corpora**

  \- **AIRBNB**(colloquial review), **POLICY**(formal news), **ParaKQC**(smart home utterances)

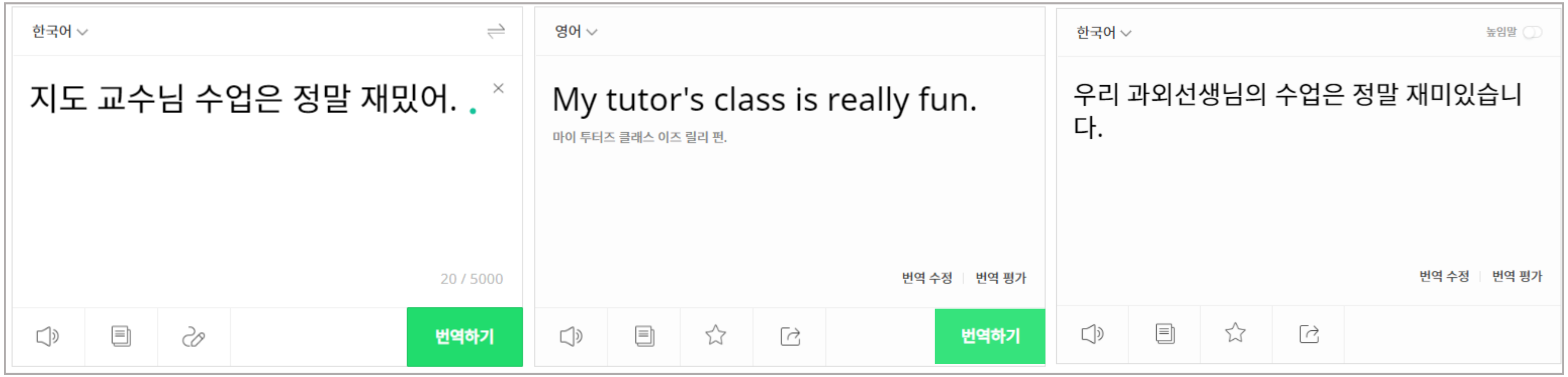
  \- **Extract or generate similar sentences and non less similar sentences**.

  Most randomized sentences similarity -> zero

	Score distribution	AIRBNB	POLICY	ParaKQC
Similar Sentences	3~5	Round-trip Translation		• Same intent as similar pairs
Less Similar Sentences	0~3	Greedy Sentence matching		• Different intent but, same topic sharing pairs

2. Semantic Textual Similarity (STS)

- **Round-trip Translation (similar sentence)**



Sentence pair	"지도 교수님 수업은 정말 재밌어." "우리 과외선생님의 수업은 정말 재미있습니다."	**Similarity : 3**

2. Semantic Textual Similarity (STS)

- **Greedy Sentence matching (less similar sentence)**

"지도 교수님 수업은 정말 재밌어."

"우리 과외 선생님 수업은 재미있습니다."

"아..수업 진짜 지루해 죽겠네."

"학교 가는 날에는 바빠서 못 봐."

"오늘 저녁 먹을 시간도 없이 바쁜데."

"수업 중인데 호박죽 먹고싶다고? "

→

"아..수업 진짜 지루해 죽겠네."

"학교 가는 날에는 바빠서 못 봐."

"오늘 저녁 먹을 시간도 없이 바쁜데."

"수업 중인데 호박죽 먹고싶다고? "

→

Sentence pair

"아..수업 진짜 지루해 죽겠네."

"수업 중인데 호박죽 먹고싶다고? "

Extract most similar sentences and remove them.

Similarity : 1

2. Semantic Textual Similarity (STS)

▪ Annotation Protocol

- 5: Two sentences are equivalent in terms of *important* and *unimportant* content.
- 4: Two sentences are closely equivalent. Some *unimportant* content differ.
- 3: Two sentences are roughly equivalent. *Important* content are similar to each other, but difference between *unimportant* content is not ignorable.
- 2: Two sentences are not equivalent. *Important* content are not similar to each other, only sharing some *unimportant* contents.
- 1: Two sentences are not equivalent. *Important* and *unimportant* content are not similar to each other. Two sentences only share their topics.
- 0: Two sentences are not equivalent. They are not sharing any *important* and *unimportant* contents and even topics.

Average 7 labels

The scores are rounded to the first decimal place.

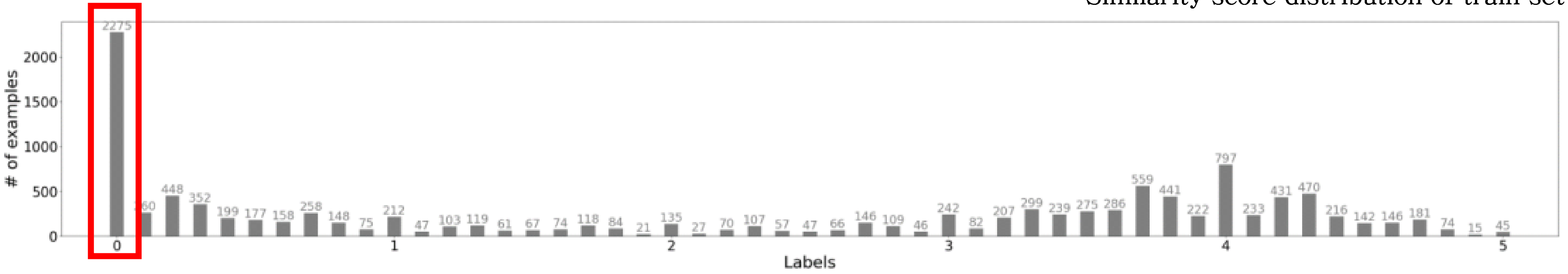
(e.g. 3.24 -> 3.2)

2. Semantic Textual Similarity (STS)

- **Final Dataset**

Source	|Train|	|Dev|	|Test|	Total
AIRBNB	5,371	255	510	6,136
POLICY	2,344	132	264	2,740
PARAKQC	3,953	132	263	4,348
Overall	**11,668**	**519**	**1,037**	**13,224**

*Similarity score distribution of train set



It is important to collect sentence pairs rigorously.

3. Natural Language Inference (NLI)

- **Why include NLI?**

  - Understanding entailment and contradiction between sentences is fundamental to NLU.
  - GLUE and superglue also include NLI.

- **Task Format**

  - Sentence-Pair Classification

- **Evaluation Metric**

  - Accuracy

3. Natural Language Inference (NLI)

- **Source Corpora (use 6 corpora)**

  - **WIKITREE**
  - **POLICY**
  - **WIKINEWS**

  → **News articles**

  - **WIKIPEDIA** → **A crowd-sourced encyclopedia**
  - **NSMC** → **Movie review(NAVER Sentiment Movie Corpus)**
  - **AIRBNB** → **Travel review**

3. Natural Language Inference (NLI)

- **Source Corpora (use 6 corpora)**

  - WIKITREE
  - POLICY
  - WIKINEWS

  → News articles

  - WIKIPEDIA → A crowd-sourced encyclopedia
  - NSMC → Movie review
  - AIRBNB → Travel review

Extract 10,000 premises! ➡

Premise should satisfy three conditions.

- A proposition, A declarative sentence.
- Must include at least one predicate.

  (be, believe, play, smile, reach -> diverse types)
- Length should be from 20 to 90 characters.

3. Natural Language Inference (NLI)

▪ What is NLI?

→ **Given a premise, an NLI model determines if hypothesis is true, false or undetermined.**

- **Entailment :** The hypothesis is necessarily true given the premise is true.
- **Contradiction :** The hypothesis is necessarily false given the premise is true.
- **Neutral :** The hypothesis may or may not be true given the premise is true.

Premise	• 고려대학교는 성북구에 위치하고 있다. (Korea University is located in Seongbuk-gu.)	
hypotheses	• 고려대학교는 서울에 위치하고 있다 (Korea University is located in Seoul.)	Entailment
	• 고려대학교는 서울 동대문구에 위치하고 있다. (Korea University is located in Dongdaemun-gu.)	Contradict
	• 고려대학교는 서울 성북구 안암동에 위치하고 있다. (Korea University is located in Anam-dong.)	Neutral

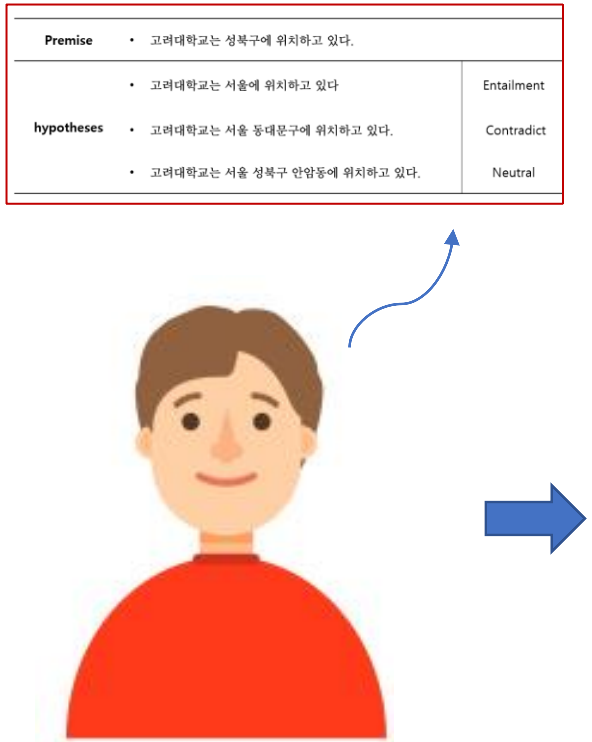
→ **Seongbuk-gu is in Seoul.**

→ **Dongdaemun-gu ≠ Seongbuk-gu**

→ **We don’t know ‘dong’ from premise.**

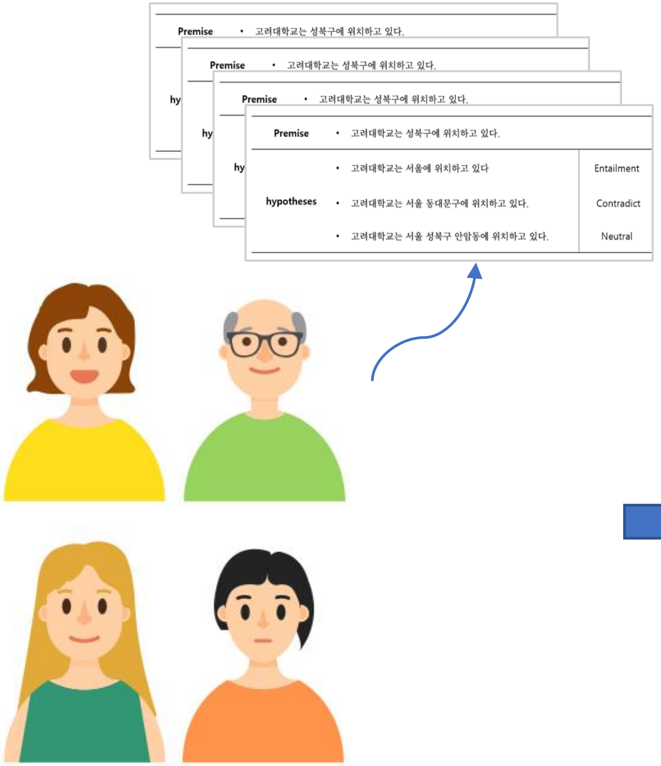
3. Natural Language Inference (NLI)

- **Annotation Protocol**



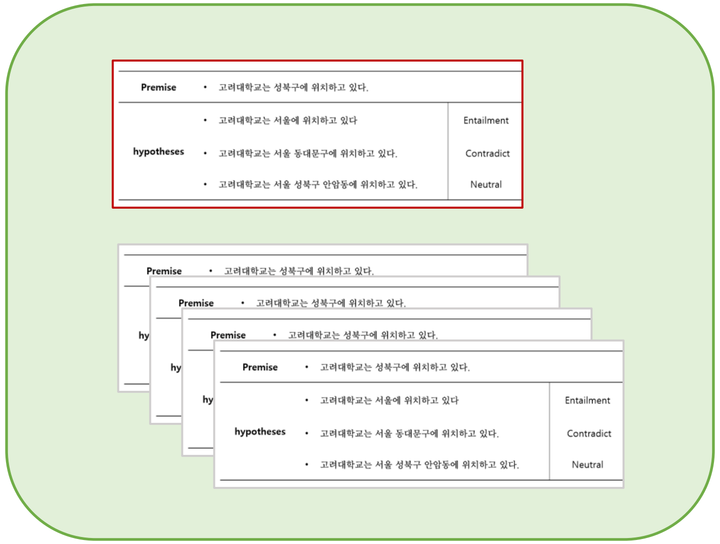
Initial Annotator

Write three hypotheses.



Four annotators

Label hypotheses.



5 labels are yielded

Gold labels

→ **Majority vote (more than 3)**

3. Natural Language Inference (NLI)

Statistics	SNLI	MNLI	KLUE-NLI
Unanimous Gold Label	58.30%	58.20%	**76.29%**
Individual Label = Gold Label	89.00%	88.70%	**92.63%**
Individual Label = Author's Label	85.80%	85.20%	**90.92%**
Gold Label = Author's Label	91.20%	92.60%	**96.76%**
Gold Label ≠ Author's Label	6.80%	5.60%	**2.71%**
No Gold Label (No 3 Labels Match)	2.00%	1.80%	**0.53%**

- **KLUE-NLI shows much higher inter-annotator agreement than SNLI and MNLI**

[1] Bowman et al., 2015, A large annotated corpus for learning natural language inference, ACL
[2] Williams et al., 2018, A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference, NAACL-HLT

3. Natural Language Inference (NLI)

▪ Human performance

Statistics	KorNLI	KLUE-NLI
Unanimous Gold Label (4 Agree)	38.00%	**71.00%**
3 Agree with Gold Label	18.00%	24.00%
2 Agree with Gold Label	18.00%	3.00%
1 Agrees with Gold Label	16.00%	2.00%
0 Agrees with Gold Label	10.00%	0.00%
Individual Label = Gold Label	64.50%	**91.00%**
No Gold Label (No 3 Labels Match)	4.00%	**0.00%**
Majority Vote ≠ Gold Label	26.00%	**0.00%**

Four native Korean undergraduates who major in Korean linguistics

- **KorNLI : translation-based Korean dataset**

- **Human performance gap provides evidence that KLUE-NLI is currently the optimal Korean NLI dataset.**

[1] Ham et al., 2020, KorNLI and KorSTS: New benchma가 datasets for Korean natural language understanding, EMNLP findings

3. Natural Language Inference (NLI)

- **Final Dataset**

Source	|Train|	|Dev|	|Test|	Total	Avg Len Prem	Avg Len Hyp
WIKITREE	3,838	450	450	4,738	52.81	26.86
POLICY	3,833	450	450	4,733	56.73	32.93
WIKINEWS	3,824	450	450	4,724	64.17	29.11
WIKIPEDIA	3,780	450	450	4,680	57.45	23.70
NSMC	4,899	600	600	6,099	27.48	21.49
AIRBNB	4,824	600	600	6,024	24.28	18.65
Overall	**24,998**	**3,000**	**3,000**	**30,998**	**47.15**	**25.46**

- 60% formal(WIKITREE/POLICY/WIKINEWS/WIKIPEDIA) and 40% colloquial(NSMC/AIRBNB) sentences.

4. Named Entity Recognition (NER)

- **Why include NER?**
  - NER is an important for application fields like syntax analysis, goal-oriented dialog system, question and answering chatbot and information extraction.
  - There are few existing Korean NER datasets.

- **Task Format**
  - Sequence Tagging → **Tagging all sequential data.**

4. Named Entity Recognition (NER)

- **Evaluation Metric**

  - Entity-level and Character-level Macro F1

  - Micro F1 score

- **Source Corpora (36,515 sentences)**

  - WIKITREE (News article corpus, formal sentences with many entity types, suitable for NER)

  - NSMC (Colloquial reviews, noisy dataset, broaden the application field of NER models)

4. Named Entity Recognition (NER)

- **What is NER?**

To detect the boundaries of named entities in unstructured text and classify the types.

##NER-1-004485 <씨엔블루:PS> 짱짱 ♥! ! ! ! <담주:DT>가 마지막이래 ㅠ. ㅠ

B-PS	**I**-PS	**I**-PS	**I**-PS	**O**	**O**	**O**	**O**	**O**	**O**	**O**	**O**	**O**	**B**-DT	**I**-DT	**O**	**O**	**O**	**O**	**O**	**O**	**O**	**O**	**O**	**O**
씨	엔	블	루		짱	짱	♥	!	!	!	!		담	주	가		마	지	막	이	래	ㅠ	.	ㅠ

(↑ An example of BIO scheme for NER tagging)

4. Named Entity Recognition (NER)

- **What is NER?**

  - **They are tagged via character-level BIO(Begin-Inside-Outside) tagging scheme.**

B- : Beginning of a chunk **I- : Inside a chunk** **O- : Token belongs to no entity/chunk**

B-PS	I-PS	I-PS	I-PS	O	O	O	O	O	O	O	O	O	B-DT	I-DT	O	O	O	O	O	O	O	O	O	O
씨	엔	블	루		짱	짱	♥	!	!	!	!		담	주	가		마	지	막	이	래	ㅠ	.	ㅠ

4. Named Entity Recognition (NER)

- **What is NER?**

  - **They are tagged via Character-leve BIO(Begin-Inside-Outside) tagging scheme.**

B- : Beginning of a chunk I- : Inside a chunk O- : Token belongs to no entity/chunk

B-PS	I-PS	I-PS	I-PS	O	O	O	O	O	O	O	O	O	B-DT	I-DT	O	O	O	O	O	O	O	O	O	O
씨	엔	블	루		짱	짱	♥	!	!	!	!		담	주	가		마	지	막	이	래	ㅠ	.	ㅠ

functional words (다음주 + -가)

+ Many compounds words in Korean contain whitespace.

e.g. "치과 의사" (dentist) → 치과의사 X

4. Named Entity Recognition (NER)

- **Entity types for KLUE-NER annotation**

- **PS(Person) : Name of individual or a group**
- **LC(Location) : Name of a district/province or a geographical location**
- **OG(Organization) : Name of an organization or an enterprise**
- **DT(Date) : Expressions related to date/period/era/age**
- **TI(Time) : Expressions related to time**
- **QT(Quantity) : Expressions related to quantity or number including units**

They follow TTA NER guidelines and MUC-7

[1] https:// committee.tta.or.kr/data/
[2] Chinchor, overview of MUC-7, ACL, 1998

4. Named Entity Recognition (NER)

- ## Annotation Process

"박보검 담주에 파리 화
보 촬영 간대! ㅠㅠ"

<박보검:PS>

<담주:DT>

<파리:LC>

Pilot Test and selecting workers

51 qualified workers annotate NER

2 linguists check the annotations

6 NLP researchers manually correct the annotation errors.

4. Named Entity Recognition (NER)

- **Annotation Protocol**

"I bought a *Cine21* from a bookstore and read it page by page."

the name of a magazine or publisher of the magazine

- **In the case of entities with multiple possible entity types, they determine their tags based on the context.**

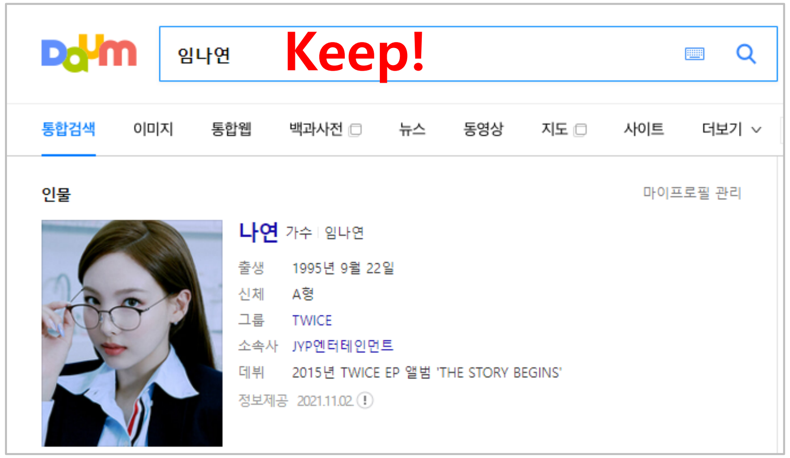
4. Named Entity Recognition (NER)

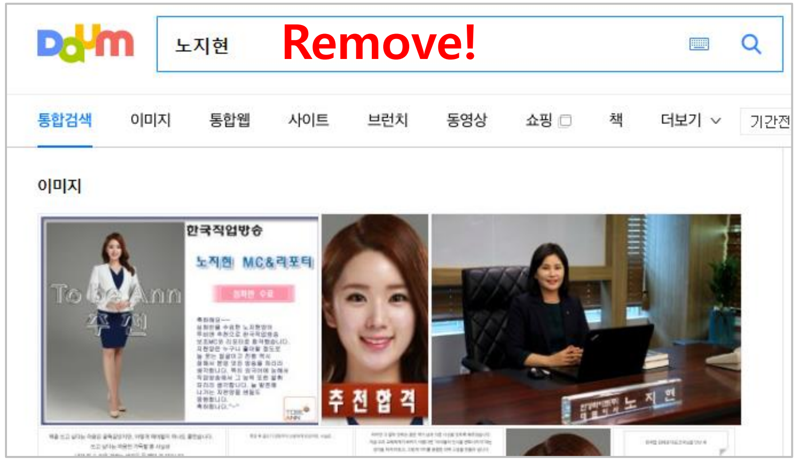
- **Annotation Protocol**

임나연 Search person name

“통화 상대는 트와이스 나연(임나연·20)씨와 쯔위 양이었다.”

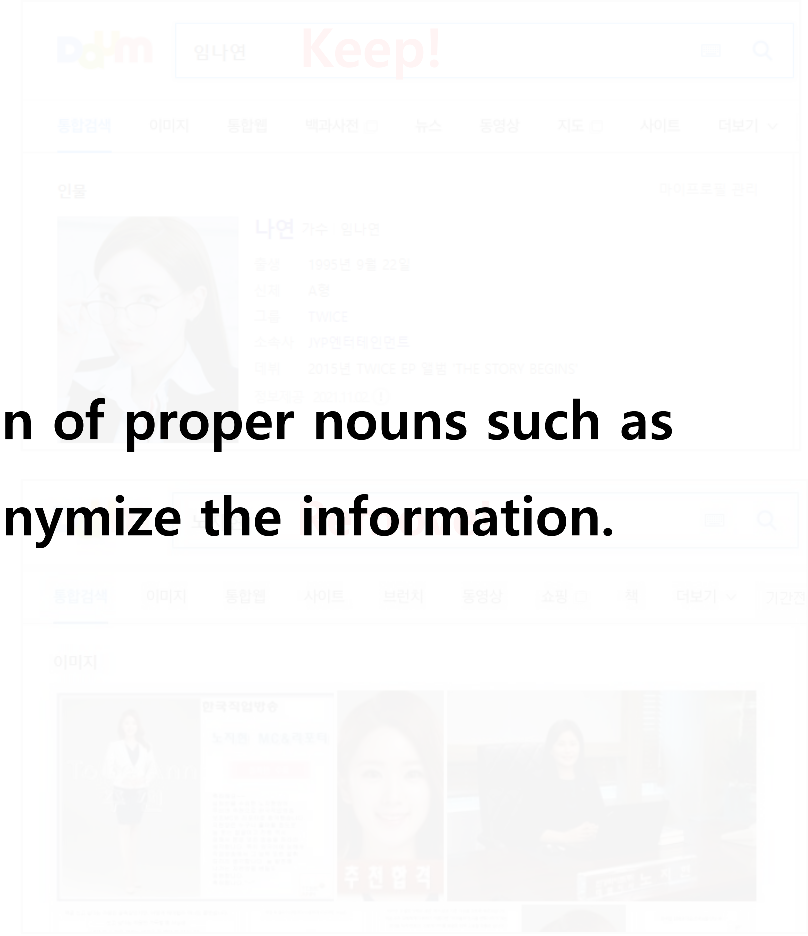
“‘느린먹거리 by 부각마을’ 대표 노지현 씨가 등장한다.”





4. Named Entity Recognition (NER)

- Annotation Protocol



In NER task often requires the specific information of proper nouns such as person name, they cannot simply drop or pseudonymize the information.

“통화 상대는 트와이스 나연(임나연·20)씨와 쯔위 양이었다.”

“‘느린먹거리 by 부각마을’ 대표 노지현 씨가 등장한다.”

4. Named Entity Recognition (NER)

- **Final dataset**

Source	|Train|	|Dev|	|Test|	Total
WIKITREE	11,435	2,534	2,685	16,664
NSMC	9,573	2,466	2,315	14,354
Total	**21,008**	**5,000**	**5,000**	$\boxed{\mathbf{31,008}}$

- **5,507 sentences are dropped in the inspection process.**

Q & A

5. Relation Extraction (RE)

- **Why include RE?**

  - RE is a task suitable for evaluating whether a model correctly understands the relationships between entities.
  - In order to ensure KLUE-RE captures this aspect of language understanding.

- **Task Format**

  - Single Sentence Classification

- **Evaluation Metric**

  - Micro F1
  - AUPRC (Area Under the Precision-Recall Curve)

5. Relation Extraction (RE)

- **What is RE?**

- RE identifies semantic relations between entity pairs in text.

- The relation is defined between an entity pair consisting of subject entity and object entity.

<키르케고르:*Subject*>는 덴마크의 수도 <코펜하겐:*Object*>의 부유한 집안에서 태어났다.

(Kierkegaard was born to an affluent family in Copenhagen.)

5. Relation Extraction (RE)

- **What is RE?**

- RE identifies semantic relations between entity pairs in text.

- The relation is defined between an entity pair consisting of subject entity and object entity.

Subject entity e_{subj} ***Object entity*** e_{obj}

⟨키르케고르 : *Subject*⟩는 덴마크의 수도 ⟨코펜하겐 : *Object*⟩의 부유한 집안에서 태어났다.

(Kierkegaard was born to an affluent family in Copenhagen.)

5. Relation Extraction (RE)

- **What is RE?**

place_of_birth(relation)

Subject entity e_{subj} *Object entity* e_{obj}

<키르케고르:*Subject*>는 덴마크의 수도 <코펜하겐:*Object*>의 부유한 집안에서 태어났다.

(Kierkegaard was born to an affluent family in Copenhagen.)

Relation triplet → $(e_{subj},\ \mathrm{r},\ e_{obj})$

5. Relation Extraction (RE)

- **Relation Schema**

Relation Class	Description
no_relation	No relation in between $(e_{\text{subj}}, e_{\text{obj}})$
org:dissolved	The date when the specified organization was dissolved
org:founded	The date when the specified organization was founded
org:place_of_headquarters	The place which the headquarters of the specified organization are located in
org:alternate_names	Alternative names called instead of the official name to refer to the specified organization
org:member_of	Organizations to which the specified organization belongs
org:members	Organizations which belong to the specified organization
org:political/religious_affiliation	Political/religious groups which the specified organization is affiliated in
org:product	Products or merchandise produced by the specified organization
org:founded_by	The person or organization that founded the specified organization
org:top_members/employees	The representative(s) or members of the specified organization
org:number_of_employees/members	The total number of members that are affiliated in the specified organization
per:date_of_birth	The date when the specified person was born
per:date_of_death	The date when the specified person died
per:place_of_birth	The place where the specified person was born
per:place_of_death	The place where the specified person died
per:place_of_residence	The place where the specified person lives
per:origin	The origins or the nationality of the specified person
per:employee_of	The organization where the specified person works
per:schools_attended	A school where the specified person attended
per:alternate_names	Alternative names called instead of the official name to refer to the specified person
per:parents	The parents of the specified person
per:children	The children of the specified person
per:siblings	The brothers and sisters of the specified person
per:spouse	The spouse(s) of the specified person
per:other_family	Family members of the specified person other than parents, children, siblings, and spouse(s)
per:colleagues	People who work together with the specified person
per:product	Products or artworks produced by the specified person
per:religion	The religion in which the specified person believes
per:title	Official or unofficial names that represent the occupational position of the specified person

Total 30 relation classes

- 18 person-related relations
- 11 organization-related relations
- no relation

5. Relation Extraction (RE)

▪ Building Process

1. Sample candidate sentences

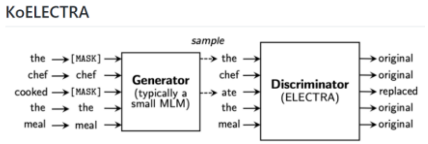
Based on TAC-KBP[1]

e_{subj} PER(Person) or ORG(Organization)

e_{obj} PER, ORG, LOC(Location), DAT(Date and time), POH(Other proper nouns), NOH(Other numerals)

2. Define Relation Schema

Fine-tune a pre-trained KoELECTRA[2] on two existing Korean NER resources [3].



3. Detect Entities

Two approaches

① KB-based sampling [4]

② Uniform sampling

(e_{subj}, e_{obj})

4. Select Entity Pairs

DeepNatural AI

• 163 qualified workers annotate each entity pair (e_{subj}, e_{obj})

5. Annotate Relations

Relation Class	Train		Dev		Test	
	Count	Ratio	Count	Ratio	Count	Ratio
Total	32,470	100.00%	7,765	100.00%	7,766	100.00%

Proportions of entity types:

PER(38.1%), ORG(36.3%), LOC(6.2%).

DAT(6.2%), POH(11.9%), NOH(1.3%)

6. Final Dataset

[1] Paul and Hoa., Overview of the TAC 2009 knowledge base population track, TAC, 2009
[2] https:// github.com/monologg/koELECTRA
[3] provided by National Institute of Korean Language and Korea Maritime & Ocean University
[4] https://aihub.or.kr/aidata/84

5. Relation Extraction (RE)

- **Final dataset**

Relation Class	Train		Dev		Test	
	Count	Ratio	Count	Ratio	Count	Ratio
Total	**32,470**	100.00%	**7,765**	100.00%	**7,766**	100.00%

<Proportions of entity types>

PER(38.1%), ORG(36.3%), LOC(6.2%).

DAT(6.2%), POH(11.9%), NOH(1.3%)

6. Dependency Parsing (DP)

- **Why include DP?**

  - DP has been an important component in NLP systems, because of its ability of capture the syntactic feature of a sentence.

- **Task Format**

  - Word-level sequence tagging task

- **Evaluation Metric**

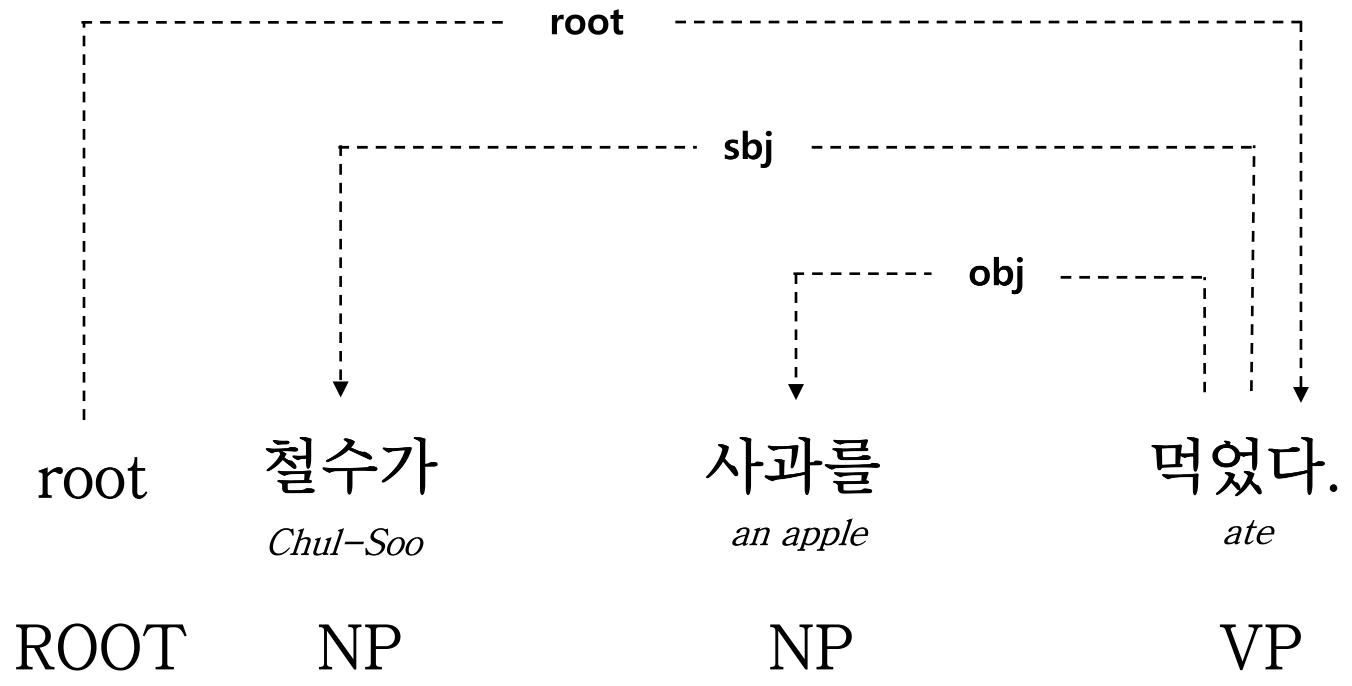
  - Unlabeled attachment score (UAS)
  - Labeled attachment score (LAS)

- **Source Corpora**

  - WIKITREE
  - AIRBNB

6. Dependency Parsing (DP)

- **What is DP?**



DP aims at finding relational information among words.

6. Dependency Parsing (DP)

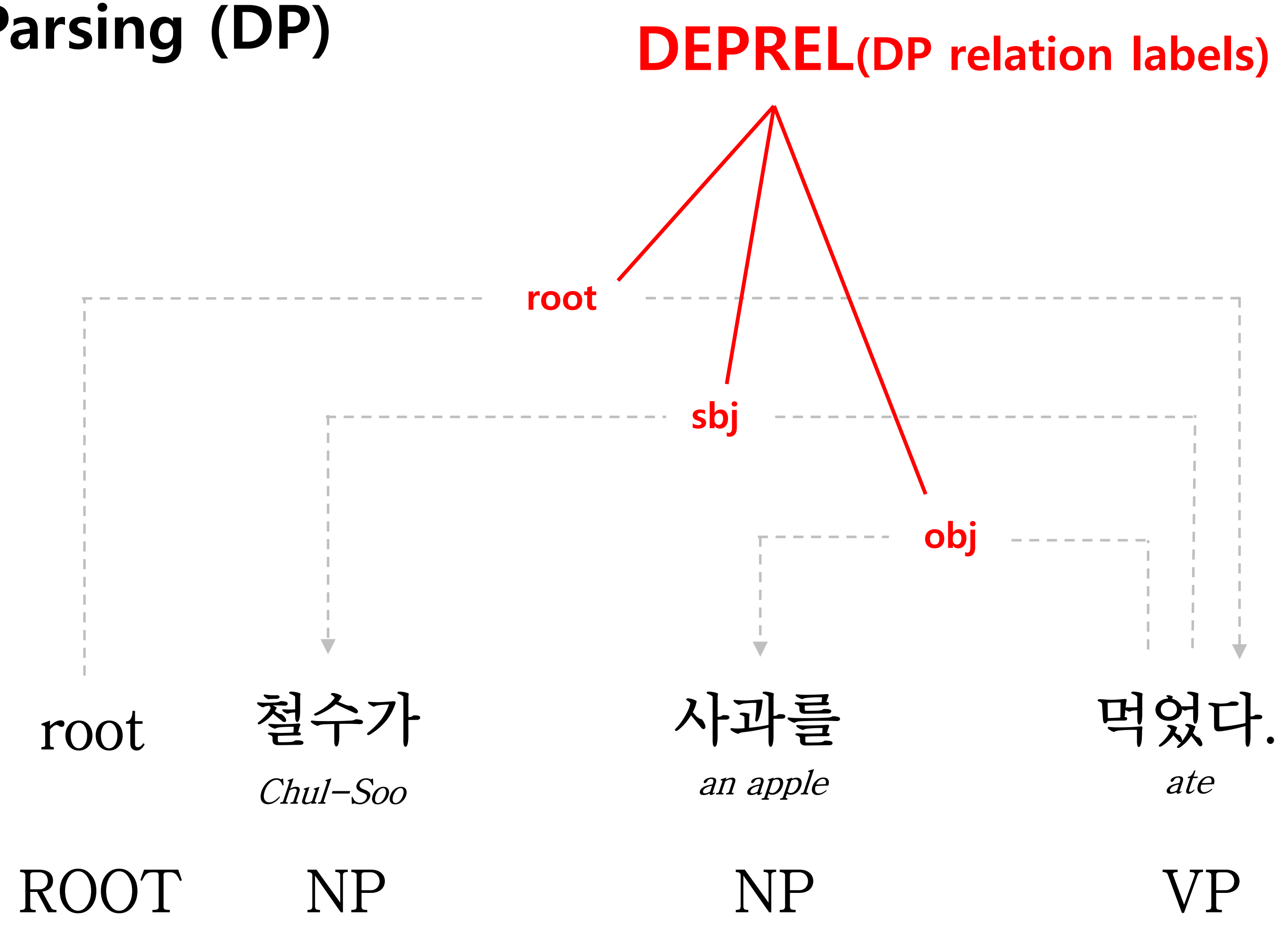
- **A demonstration of the output format**

##DP−ex−000000 철수가 사과를 먹었다. (*Chul−Soo* ate an apple.)

Word index	Word Form	Tokenization	POS tag	Head of the current word	DEPREL
1	철수가	철수 가	NNP + JKS	3	NP_SBJ
2	사과를	사과 를	NNG + JKO	3	NP_OBJ
3	먹었다.	먹 었 다.	VV + EP + EF + SF	0	VP

6. Dependency Parsing (DP)

- **What is DP?**



DP aims at finding relational information among words.

6. Dependency Parsing (DP)

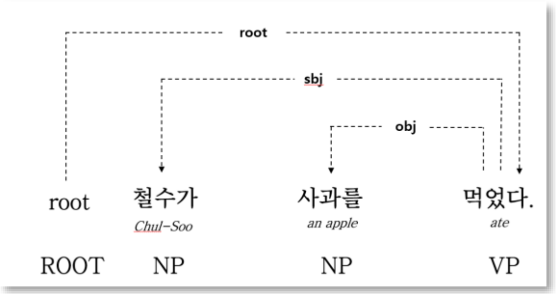
- ## Annotation Protocol and Process

1. POS(part-of-speech) annotation

철수 가	NNP + JKS
사과 를	NNG + JKO
먹 었 다.	VV + EP + EF + SF

- To utilize POS information as an additional syntactic feature.

2. Dependency Relation annotation



NP_SBJ

NP_OBJ

VP

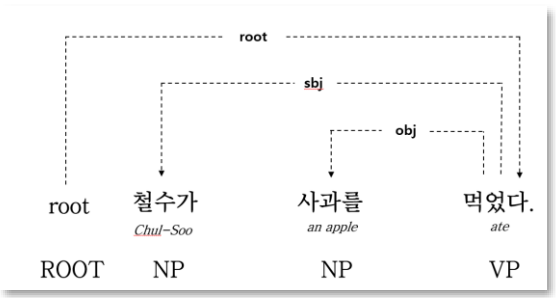
- They modify the original TTA DP guideline for dependency relation annotation. (add guides for spoken and web data)

6. Dependency Parsing (DP)

▪ Annotation Protocol and Process

Label Type	Description
Syntax	
NP	Noun Phrase
VP	Verb Phrase
AP	Adverb Phrase
VNP	Copula Phrase
DP	Adnoun Phrase
IP	Interjection Phrase
X	Pseudo Phrase
L	Left Parenthesis and Quotation Mark
R	Right Parenthesis and Quotation Mark
Function	
SBJ	Subject
OBJ	Object
MOD	Noun Modifier
AJT	Predicate Modifier
CMP	Complement
CNJ	Conjunction

2. Dependency Relation annotation



NP_SBJ

NP_OBJ

VP

- **They modify the original TTA DP guideline for dependency relation annotation. (add guides for spoken and web data)**

6. Dependency Parsing (DP)

- **Annotator**

10 PhD students in Korean linguistics

- **Final Dataset**

Source	|Train|	|Dev|	|Test|	Total
WIKITREE	5,000	1,000	1,250	7,250
AIRBNB	5,000	1,000	1,250	7,250
Total	**10,000**	**2,000**	**2,500**	**14,500**

7. Machine Reading Comprehension(MRC)

- **Why include MRC?**

  - In Korean, an appropriate MRC benchmark is not available.
  - To construct challenging Korean MRC dataset.

- **Task Format** : Answer span prediction **(Span is a slice from the document.)**

  - **Q** : 오바마는 언제부터 대통령 임기를 시작했어? (When did Obama start his presidency?)
  - **P** : … 오바마는 하와이 출신으로 대통령에 당선 되었으며, **2009년 1월 20일**부터 임기를 시작하였다…
    (…. Obama was elected president from Hawaii and began his term on January 20, 2009…..)
  - **A** : **2009년 1월 20일**. (January 20, 2009)

- **Evaluation Metric**

  - ROUGE-W(Longest common consecutive subsequence based)
  - Exact Match(EM)

7. Machine Reading Comprehension(MRC)

- **Source Corpora**

  - WIKIPEDIA
  - The Korea Economy Daily
  - ACROFAN

- **Contributions**

  - Providing multiple question types.
  - Preventing reasoning shortcuts.
  - Multiple passage domains accessible to everyone. (To guarantee CC BY-SA license)

7. Machine Reading Comprehension(MRC)

- **What is MRC?**

"올여름 장마가 17일 제주도에서 시작됐다. 서울 등 중부지방은 예년보다 사나흘 정도 늦은 이달 말께 장마가 시작될 전망이다. 17일 기상청에 따르면 제주도 남쪽 먼바다에 있는 장마전선의 영향으로 이날 제주도 산간 및 내륙지역에 호우주의보가 내려지면서 곳곳에 100mm에 육박하는 많은 비가 내렸다. 제주의 장마는 평년보다 2~3일, 지난해보다는 하루 일찍 시작됐다. 장마는 고온다습한 북태평양 기단과 한랭 습윤한 오호츠크해 기단이 만나 형성되는 장마전선에서 내리는 비를 뜻한다. 장마전선은 18일 제주도 먼 남쪽 해상으로 내려갔다가 20일께 다시 북상해 전남 남해안까지 영향을 줄 것으로 보인다. 이에 따라 20~21일 남부지방에도 예년보다. 사흘 정도 장마가 일찍 찾아올 전망이다. 그러나 장마전선을 밀어올리는 북태평양 고기압 세력이 약해 서울 등 중부지방은 평년보다 사나흘가량 늦은 이달 말부터 장마가 시작될 것이라는 게 기상청의 설명이다. 장마전선은 이후 한 달가량 한반도 중남부를 오르내리며 곳곳에 비를 뿌릴 전망이다. 최근 30년간 평균치에 따르면 중부지방의 장마 시작일은 6월24~25일이었으며 장마기간은 32일, 강수일수는 17.2일이었다. 기상청은 올해 장마기간의 평균 강수량이 350~400mm로 평년과 비슷하거나 적을 것으로 내다봤다. 브라질 월드컵 한국과 러시아의 경기가 열리는 18일 오전 서울은 대체로 구름이 많이 끼지만 비는 오지 않을 것으로 예상돼 거리 응원에는 지장이 없을 전망이다.",

▪ **Question**	"북태평양 기단과 오호츠크해 기단이 만나 국내에 머무는 기간은? "
▪ **Answer**	"한 달가량", "한 달"

7. Machine Reading Comprehension(MRC)

▪ Annotation Protocol

… 오바마는 하와이 출신으로 대통령에 당선 되었으며, 2009년 1월 20일부터 임기를 시작하였다…

Q. 오바마는 언제부터 대통령 임기를 시작했어?

A. 2009년 1월 20일.

- **Question types**

- **Type1** Question Paraphrasing
- **Type2** Multiple-Sentence Reasoning
- **Type3** Unanswerable Questions

Workers generate questions and label answers spans

7. Machine Reading Comprehension(MRC)

- **Annotation Process**

  - **Crowd workers**

	Annotator	Inspector
Type 1	28	3
Type 2	19	3
Type 3	13	2

✓ **If the generated question is rejected by the inspectors, it it regenerated.**

✓ **Through the filtering process, they remove 173 examples in total.**

→ **manually re-check all examples at the end of the annotation process**

7. Machine Reading Comprehension(MRC)

- **Final Dataset**

  - 12,207 paraphrasing-based questions.
  - 7,895 multi-sentence reasoning questions.
  - 9,211 unanswerable questions.

Total 29,313 examples

22,343 documents and 23,717 passages

	|Train|	|Dev|	|Test|	Total
# Documents	12,174	5,075	5,094	22,343
# Passages	13,072	5,310	5,335	23,717
# Questions	17,554	5,841	5,918	29,313
Avg Length of Passage	1,004.62	1,014.64	1,010.13	1,008.10
Avg Length of Question	29.00	29.05	29.01	29.01
Avg Length of Answer	6.03	6.03	5.82	5.99

6:2:2

(train/dev/test split ratio)

8. Dialogue State Tracking (DST, Wizard-of-Seoul)

- **Why include DST?**
  - Building a human-computer conversation system has been increasingly attracting attention.
  - DST is a core module of task-oriented dialogue systems.

- **Task Format**
  - Slot-value Prediction

- **Evaluation Metric**
  - JGA : Joint goal accuracy.
  - Slot micro F1 score.

- **Source Corpora**
  - Construct via *Self-Dialog*

8. Dialogue State Tracking (DST, Wizard-of-Seoul)

- DST is predicting the dialogue states from a given dialogue context.

user 안녕하세요.

DS : [ ]

Dialogue states

sys 네. 안녕하세요. 무엇을 도와드릴까요?

user 서울 중앙에 위치한 호텔을 찾고 있습니다. 외국인 친구도 함께 갈 예정이라서 원활하게 인터넷을 사용할 수 있는 곳이었으면 좋겠어요.

DS : [(숙소-지역, 서울 중앙), (숙소-종류, 호텔), (숙소-인터넷 가능, yes)] → Slot and value pair

sys 네. 확인해보겠습니다. 혹시 추가로 필요하신 사항이 있으실까요?

user 음.. 예약 인원은 총 8명 이고요. 아. 가격대는 크게 상관 없습니다.

DS : [(숙소-지역, 서울 중앙), (숙소-종류, 호텔), (숙소-인터넷 가능, yes), (숙소-예약 명수, 8), (숙소-가격대, dontcare)]

sys 네. 확인 감사합니다. 숙박을 원하시는 요일과 기간 같이 확인 부탁드립니다.

user 아. 중요한 걸 깜빡했네요. 일요일에 2일간 예약하고 싶습니다.

DS : [(숙소-지역, 서울 중앙), (숙소-종류, 호텔), (숙소-인터넷 가능, yes), (숙소-예약 명수, 8), (숙소-가격대, dontcare), (숙소-예약 요일, 일요일), (숙소-예약기간, 2)]

8. Dialogue State Tracking (DST, Wizard-of-Seoul)

▪ Building Wizard-of-Seoul (WoS)

1. Define task schema

Domains	Informable Slots	Requestable Slots
Hotel	name, type*, area*, price range*, book day†, book stay†, book people†, walkability*, parking*, internet*, breakfast*, smoking*, fitness*, swimming pool*, spa*	rating, nearby station, minutes walk from station, address, phone number, business hour, reference number‡
Restaurant	name, type*, area*, price range*, book day†, book time†, book people†, alcohol*, walkability*, parking*, internet*, smoking*, outdoor table*	rating, nearby station, minutes walk from station, address, phone number, business hour, last order time, representative menu, reference number‡
Attraction	name, type*, area*, walkability*, parking*, heritage*, educational*, scenic*, cultural*	rating, nearby station, minutes walk from station, address, phone number, business hour, entrance fee
Taxi	leave at*, departure*, arrive by, destination*, type	phone number, cost, duration
Metro	leave at, departure*, destination*	departure line, destination line, arrive by, cost, duration, transfer, optimal path

2. Create Knowledge base

Domain	# Instances	# Slots
Hotel	101	19
Restaurant	56	20
Attraction	100	17
Taxi	-	8
Metro	3,306	10

3. Design an annotation system

Korean

당신은 오늘 **22:41**에 **서울 중앙**에서 식사할 계획을 가지고 있습니다. 아참 오늘은 **수요일** 입니다. 그런 곳을 찾았다면 먼저 **대표 메뉴**를 확인하세요. 그리고 나선 **1**명으로 예약 거세요. 예약 이후엔 **영업 시간**을 문의하시구요. 그리고 나선 **식당 근처**에서 잘 곳을 찾아야 합니다. 그 곳은 반드시 **흡연이 불가**해야 합니다. 찾았다면 **같은 요일**에 예약하세요. **같은 인원**으로 **4**일간 머물러야 합니다. 예약에 성공했다면 **예약 번호**를 묻고, **흡연 가능 유무**를 더블 체크하세요. 그런 다음 마지막으로 택시를 하나 부르세요. **식당**에서 **숙소**로 향해야 합니다. 찾았다면 **소요 시간**을 문의하세요.

4. Collect and annotate a dataset

Domains, informable slots, requestable slots

Construct a KB based on task schema of each domains

Provide a goal instruction

They adapt 'Self-dialog' scheme(less costive

8. Dialogue State Tracking (DST, Wizard-of-Seoul)

- **Final Dataset**

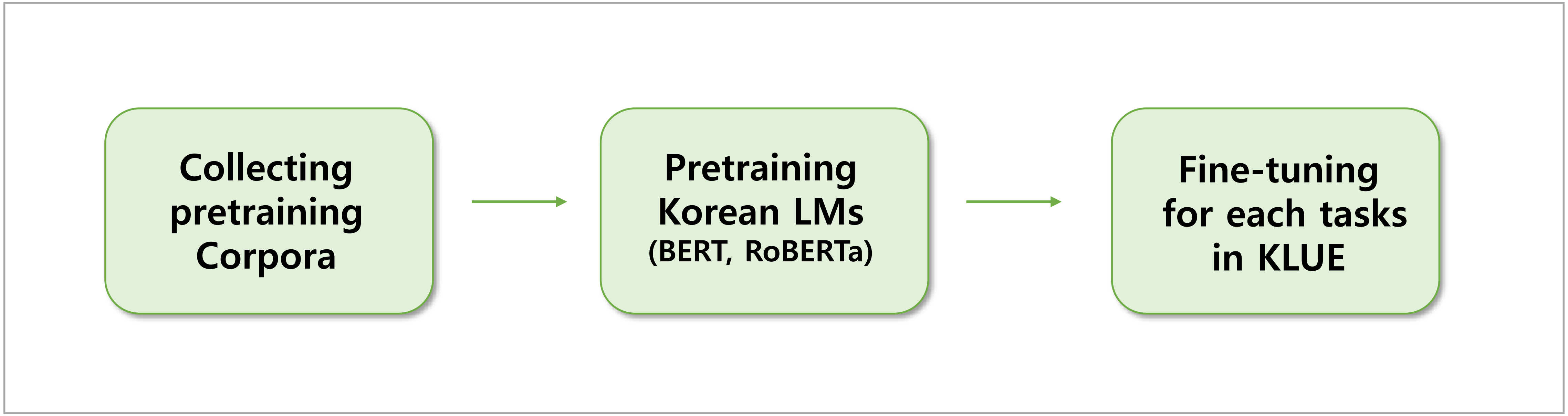
	|Train|	|Dev|	|Test|	Total
# Dialogues	8,000	1,000	1,000	10,000
# Single Domain Dialogues	1,806	263	226	2,295
# Multi Domain Dialogues	6,194	737	774	7,705
# Counterfactual Dialogues	0	294	361	655
# Total Turns	117,584	14,448	14,660	146,692
# Total Tokens	899,450	114,169	114,914	1,128,533
Avg Turns per Dialogue	14.70	14.45	14.66	14.67
Avg Tokens per Turn	7.65	7.90	7.84	7.69

One dialogue
-> 14.67 turns

WoS contains overall 10,000 dialogues with 146,692 turns across 5 domains.

Experiments

Pretrained Language Models



They pretrain and release large-scale LMs for Korean.

Pretrained Language Models

- Pretrainig Corpora

	MODU	CC-100-Kor	NAMUWIKI	NEWSCRAWL	PETITION	Total
# Sentences	167M	103M	14M	183M	5.2M	**473M**
# Words	1,892,814,395	1,593,887,022	265,203,602	2,716,968,038	50,631,183	**6,519,504,240**
size (GB)	18.27	15.46	2.52	25.87	0.53	**62.65**

- Tokenization

Tokenization	Tokenized Sequence
Raw Text	조경현은 인공지능 분야의 저명한 연구자이다.
BPE (Multilingual)	조 / ##경 / ##현 / ##은 / 인 / ##공 / ##지 / ##능 / 분 / ##야 / ##의 / 저 / ##명한 / 연구 / ##자 / ##이다 / .
BPE	조경 / ##현은 / 인공지능 / 분야의 / 저 / ##명한 / 연구 / ##자이 / ##다 / .
Morpheme	조경현 / 은 / 인공지능 / 분야 / 의 / 저명 / 한 / 연구자 / 이 / 다 / .
Morpheme-based Subword	조경 / ##현 / ##은 / 인공지능 / 분야 / ##의 / 저명 / ##한 / 연구자 / ##이다 / .

[1] corpus.Korean.go.kr
[2] Data.statmt.org/cc-100/
[3] dump.thewiki.kr
[4] www1.president.go.kr/petitions

Pretrained Language Models

- **Pretrained LMs Experiment Settings**

Model	# Parameter	Masking	Training Steps	Batch Size	Learning Rate	Device
KLUE-BERT$_{\text{BASE}}$	110M	Static, WWM	1M	256	10^{-4}	TPU v3-8
KLUE-RoBERTa$_{\text{SMALL}}$	68M	Dynamic, WWM	1M	2048	10^{-4}	8× V100 GPUs
KLUE-RoBERTa$_{\text{BASE}}$	110M	Dynamic, WWM	1M	2048	10^{-4}	8× V100 GPUs
KLUE-RoBERTa$_{\text{LARGE}}$	337M	Dynamic, WWM	500k	2048	10^{-4}	8× V100 GPUs

- **Baseline models**
  - mBERT(multilingual BERT)
  - XLM-R (multilingual RoBERTa)
  - KR-BERT (Korean BERT)
  - KoELECTRA (Korean ELECTRA)

Pretrained Language Models

- **Pseudonymization**

Private Information	Pseudonymization	Pseudonymised Example
Telephone Number	Faker	055-604-8764
Social Security Number	Faker	600408-2764759
Foreign Registration Number	Faker	110527-1815659
Email Address	Faker	agweon@example.org
IP Address	Faker	166.186.169.69
MAC Address	Faker	c5:d7:14:84:f8:cf
Mention(@)	Faker	@gildong
Address	Random Number Generation	경상북도 성남시 서초대64가
Bank Account Number	Random Number Generation	110-245-124678
Passport Number	Random Generation	M123A4567
Driver's License	Random Number Generation	11-17-174133-01
Business Registration Number	Random Number Generation	123-45-67890
Health Insurance Information	Random Number Generation	1-2345678901
Credit or Debit Card Number	Random Number Generation	1234-5678-9012-3456
Vehicle Registration Place	Random Generation	55구 1601
Homepage URL	Random Generation	www.example.com

✓ They pseudonymize PII(personally identifiable information) to avoid ethical issue.

✓ They detect 16 personal data types.

✓ 1.2% of the pretraining corpora.

[1] https://faker.readthedocs.ip

Fine-tuning Language Models

- **Task-Specific Architectures**

→ 8 tasks can be categorized into 4 types based on the fine-tuning strategy.

Single sentence classification	Sentence pair classification / regression	Multiple-sentence Slot-Value Prediction	Sequence Tagging
↓	↓	↓	↓
Topic Classification Relation Extraction	Sentence Textual Similarity Natural Language Inference	Dialogue state tracking	Named Entity Recognition Dependency Parsing Machine Reading Comprehension

Fine-tuning Language Models

	YNAT	KLUE-STS		KLUE-NLI	KLUE-NER		KLUE-RE		KLUE-DP		KLUE-MRC		WoS	
Model	F1	R^P	F1	ACC	$F1^E$	$F1^C$	$F1^{mic}$	AUC	UAS	LAS	EM	ROUGE	JGA	$F1^S$
mBERT$_{BASE}$	81.55	84.66	76.00	73.20	76.50	89.23	57.88	53.82	90.30	86.66	44.66	55.92	35.46	88.63
XLM-R$_{BASE}$	83.52	89.16	82.01	77.33	80.37	92.12	57.46	54.98	89.20	87.69	27.48	53.93	39.82	89.61
XLM-R$_{LARGE}$	**86.06**	92.97	85.86	85.93	82.27	**93.22**	58.39	61.15	92.71	**88.70**	35.99	66.77	41.20	89.80
KR-BERT$_{BASE}$	84.58	88.61	81.07	77.17	74.58	90.13	62.74	60.94	89.92	87.48	48.28	58.54	45.33	90.70
KoELECTRA$_{BASE}$	84.59	92.46	84.84	85.63	**86.11**	92.56	62.85	58.94	92.90	87.77	59.82	66.05	41.58	89.60
KLUE-BERT$_{BASE}$	85.73	90.85	82.84	81.63	83.97	91.39	66.44	66.17	89.96	88.05	62.32	68.51	46.64	91.61
KLUE-RoBERTa$_{SMALL}$	84.98	91.54	85.16	79.33	83.65	91.14	60.89	58.96	90.04	88.14	57.32	62.70	46.62	91.44
KLUE-RoBERTa$_{BASE}$	85.07	92.50	85.40	84.83	84.60	91.44	67.65	68.55	93.04	88.32	68.67	73.98	47.49	91.64
KLUE-RoBERTa$_{LARGE}$	85.69	**93.35**	**86.63**	**89.17**	85.00	91.86	**71.13**	**72.98**	**93.48**	88.36	**75.58**	**80.59**	**50.22**	**92.23**

KLUE-RoBERTa LARGE model performs best on several tasks.

Fine-tuning Language Models

	YNAT	KLUE-STS		KLUE-NLI	KLUE-NER		KLUE-RE		KLUE-DP		KLUE-MRC		WoS	
Tokenization	F1	R^P	F1	ACC	$F1^E$	$F1^C$	$F1^{mic}$	AUC	UAS	LAS	EM	ROUGE	JGA	$F1^S$
BPE	**83.40**	91.91	**85.19**	**82.07**	68.75	89.47	64.39	**65.04**	89.89	**89.47**	51.12	65.79	21.38	77.68
Morpheme-based Subword	**83.40**	**92.06**	84.70	81.60	**84.84**	**91.03**	**65.25**	64.79	**92.17**	88.34	**62.13**	**67.46**	**47.14**	**91.60**

KLUE Leaderboard

Unlike other benchmarks, klue benchmarks do not provide total scores and leaderboards for the entire task. On the leaderboard, you can check each score for one model and sort by each evaluation metric.

All | Small Size | Base Size | Large Size

#	Team	Model	Description	YNAT	KLUE-STS		KLUE-NLI	KLUE-NER		KLUE-RE		KLUE-DP		KLUE-MRC		WOS	
				F1	R^P	F1	ACC	$F1^E$	$F1^C$	$F1^{mic}$	AUC	UAS	LAS	EM	ROUGE	JGA	$F1^S$
1	KLUE-team	KLUE-BERT-base	More	85.73	90.85	82.84	81.63	83.97	91.39	66.44	66.17	89.96	88.05	62.32	68.51	46.64	91.61
2	KLUE-team	KLUE-RoBERTa-large	More	85.69	93.35	86.63	89.17	85	91.86	71.13	72.98	93.48	88.36	75.58	80.59	50.22	92.23
3	KLUE-team	KLUE-RoBERTa-base	More	85.07	92.5	85.4	84.83	84.6	91.44	67.65	68.55	93.04	88.32	68.67	73.98	47.49	91.64
4	KLUE-team	KLUE-RoBERTa-small	More	84.98	91.54	85.16	79.33	83.65	91.14	60.89	58.96	90.04	88.14	57.32	62.7	46.62	91.44
5	KLUE-tester		More	79.63	88.51	81.22	67.03	81.07	89.39	44.86	31.99	89.58	88.03	40.74	45.86	2.44	48.04

Rows per page: 5 1-5 of 5

Conclusion

- **They present KLUE, a suite of Korean NLU benchmark that includes diverse tasks.**
- **They provide pretrained large-scale language models for Korean.**
- **Their benchmark KLUE will facilitate future Korean NLP research.**

Q & A