KLUE

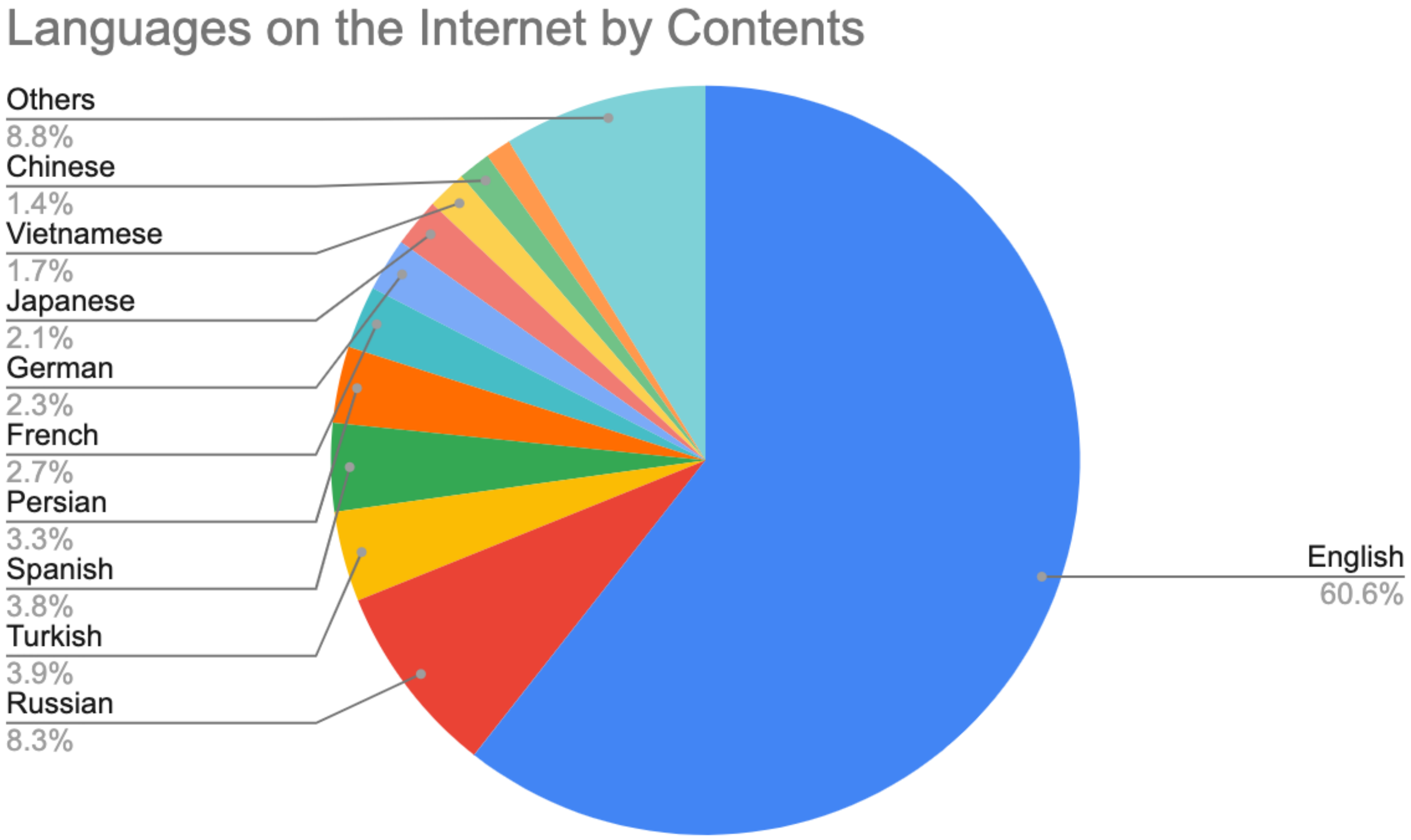
Korean Language Understanding Evaluation

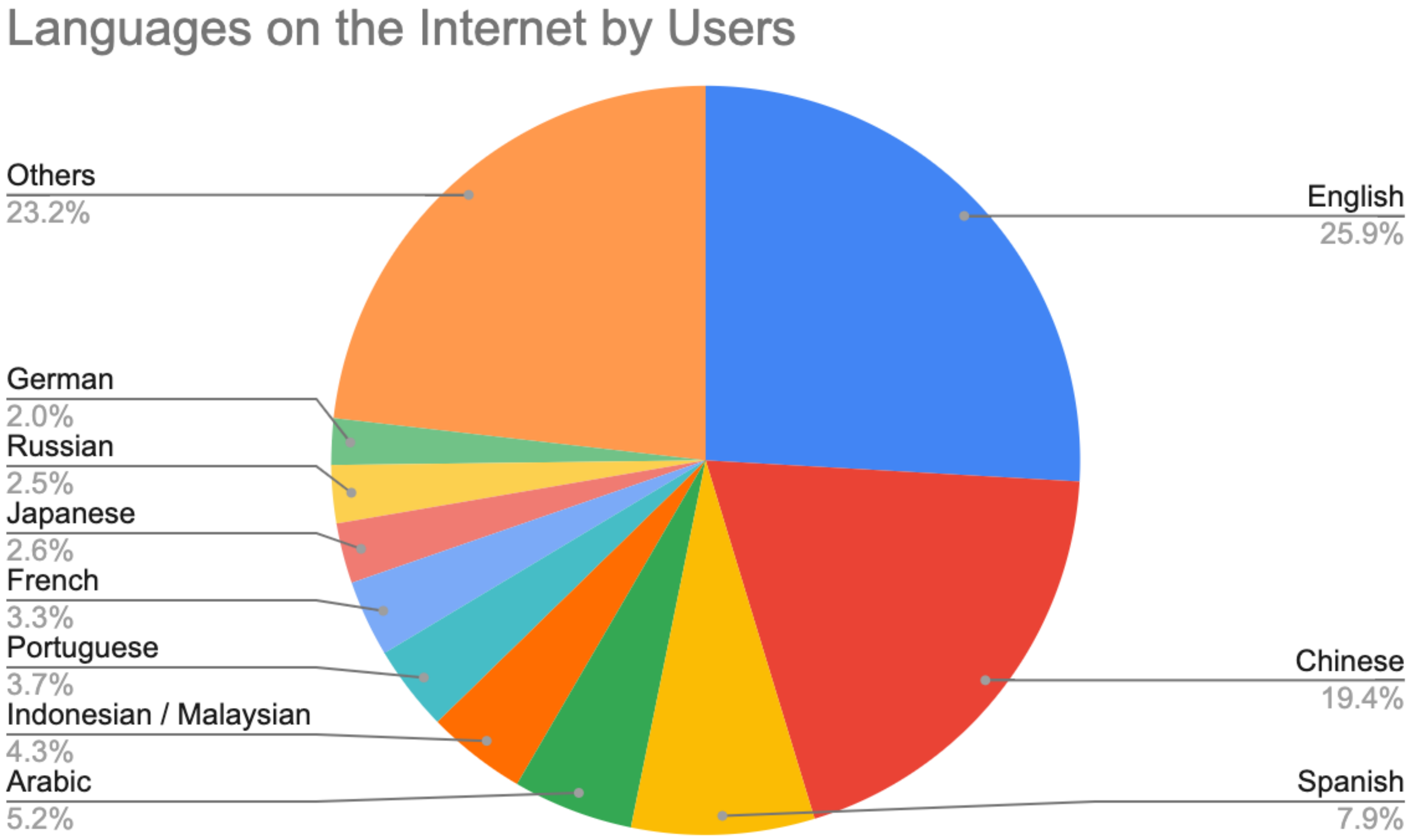
Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, Kyunghyun Cho

Building a benchmark suite for a new language

A growing discrepancy between language users and content availability

- https://en.wikipedia.org/wiki/Languages_used_on_the_Internet
- Not too different when we look at the resource availability for NLP





Building a benchmark suite for a new language

Translating existing corpora into a new language

- Machine translation has advanced greatly
  - Automatically translate training instances
- Professional translation is pretty much perfect
  - Manually translate validation/test instances
- XNLI [Conneau et al., 2018] is a representative example
  - extends MNLI [Williams et al., 2017] into 15 languages by professional translation

Building a benchmark suite for a new language

Translation may be enough

- **Translating** an original corpus **into a new languages**
  - **Advantages**
    - Minimal discrepancy between the original and new corpora
    - Instance-level comparison between two languages is possible
    - The strength of the original corpus transfers to the new corpus

Building a benchmark suite for a new language

Translation is not enough

- **Translating** an original corpus **into a new languages**
  - **Disadvantages**
    - Cultural/social discrepancy between the original and target languages
    - Translationese vs. natural language
      - Wintner's tutorial at COLING'16
        <Translationese: between human and machine translation>
    - The weakness of the original corpus transfers to the new corpus

Building a benchmark suite for a new language

Building it from scratch

- We can build a benchmark suite for a new language **from scratch**.
- **Advantages**
    - (Fairly) accurately reflects **social/cultural norms** of target-language speakers.
    - Can use the **best practices** of data construction known so far.

Building a benchmark suite for a new language

Building it from scratch

- We can build a benchmark suite for a new language **from scratch**.
- **Advantages**
    - (Fairly) accurately reflects **social/cultural norms** of target-language speakers.
    - Can use the **best practices** of data construction known so far.
- **Disadvantages**
    - **Capital intensive**: purchasing source corpora, manual annotation, etc.
    - **Labor intensive**: Manual annotation, manual quality control, etc.

Building a benchmark suite for Korean

from scratch

- **Korean**
  - More than **75M** (native) **speakers**
    - Mostly in South Korea, North Korea and a part of China.
  - Language isolate
    - Koreanic - Korean
  - Writing system
    - Hangul 한글

조선글

한글

Building a benchmark suite for Korean

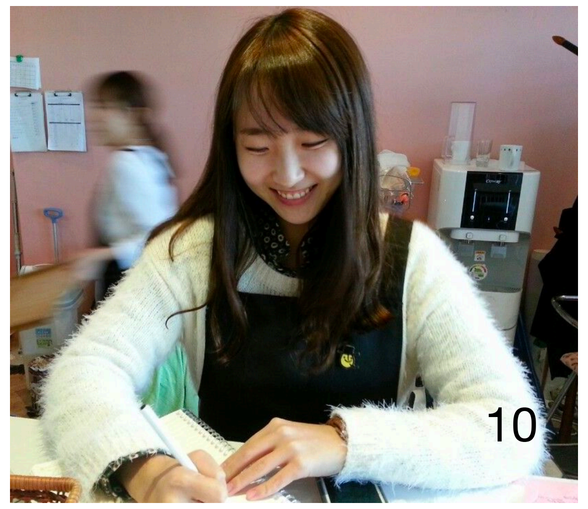
from scratch

- **Korean**
  - More than **75M** (native) **speakers**
  - Language isolate
  - Writing system: 한글
- **No** benchmark suite for evaluating Korean language understanding systems
  - A few benchmark datasets scattered here and there.

Building a benchmark suite for Korean

Korean Language Understanding Evaluation (KLUE)

- **30+ researchers** from **12 organizations** in Korea
- **11 sponsors**
  - Financial sponsors
  - Compute sponsors
  - Data sponsors
- Led by **Sungjoon Park** and **Jihyung Moon** (both Upstage.AI)



Four principles

KLUE

- **Accessibility**
  - KLUE must be openly usable by all, including academia and industry
  - KLUE must facilitate future advances: must allow derivatives.
- **Diversity**
- **Accurate annotation**
- **Safety**

Four principles

KLUE

- **Accessibility**
- **Diversity**
  - KLUE must cover diverse aspects of language understanding
  - KLUE must cover diverse topics and styles
- **Accurate annotation**
- **Safety**

Four principles
KLUE

- **Accessibility**
- **Diversity**
- **Accurate annotation**
  - KLUE must provide annotations that are accurate and unambiguous
- **Safety**

Four principles
KLUE

- **Accessibility**
- **Diversity**
- **Accurate annotation**
- **Safety**
  - KLUE must proactively deal with social biases and toxic contents

Let’s build KLUE

Task selection

Classification

- Topic classification
- Semantic textual similarity
- Natural language inference
- Named entity recognition
- Relation extraction
- Dependency parsing
- Machine reading comprehension
- Dialogue state tracking

Single-sentence classification

checks the ability of capturing the semantics of text

Task selection

Classification

- Topic classification
- Semantic textual similarity
- Natural language inference
- Named entity recognition
- Relation extraction
- Dependency parsing
- Machine reading comprehension
- Dialogue state tracking

Multi-sentence classification

checks the ability of capturing relationship among multiple sentences

Task selection

Structured prediction

- Topic classification
- Semantic textual similarity
- Natural language inference
- Named entity recognition
- Relation extraction
- Dependency parsing
- Machine reading comprehension
- Dialogue state tracking

Tagging

checks the ability of identifying important portions of text in the context of a target task or a given context

Task selection

Structured prediction

- Topic classification
- Semantic textual similarity
- Natural language inference
- Named entity recognition
- Relation extraction
- Dependency parsing
- Machine reading comprehension
- Dialogue state tracking

Graph induction (advanced tagging)

checks the ability of capturing the relationship among the words within text

Task selection

Structured prediction

- Topic classification
- Semantic textual similarity
- Natural language inference
- Named entity recognition
- Relation extraction
- Dependency parsing
- Machine reading comprehension
- Dialogue state tracking

Slot filling (advanced tagging)

checks the ability of capturing the relationship across multiple utterances in the context of information collection

Task selection

Generation is left for the future

- Although **generation** is a key aspect of language understanding, there are a number of **challenges**:
  - **Evaluation**: how do we properly evaluate the quality of generated text?
  - **Annotation**: how do we collect a diverse set of text per instance to properly?
- We thus leave out generation for now.

- Topic classification
- Semantic textual similarity
- Natural language inference
- Named entity recognition
- Relation extraction
- Dependency parsing
- Machine reading comprehension
- Dialogue state tracking

Task selection

Review: criteria

- **Diversity**
  - Diverse aspects of language understanding
  - Diverse task formats
- **Evaluation**
  - (Somewhat) objective evaluation metrics exist
- **Annotation**
  - Unambiguous targets (often) exist

- Topic classification
- Semantic textual similarity
- Natural language inference
- Named entity recognition
- Relation extraction
- Dependency parsing
- Machine reading comprehension
- Dialogue state tracking

Curating source corpora

Considerations

- For each potential source corpus, we consider
  - **License**
  - **Domain**
  - **Style**: formal vs. colloquial, modern vs. not
  - **Ethical risks**
  - **Volume/Size**

Let's look at a few examples!

Curating source corpora

News Headlines

- **License**: N/A, because these are only headlines
- **Domain**: News
- **Style**: formal, modern
- **Ethical risks**: low
- **Volume/Size**: large
- **INCLUDED!**

Curating source corpora

National Assembly Minutes

- **License**: public domain
- **Domain**: politics
- **Style**: colloquial, modern
- **Ethical risks**: medium
- **Volume/Size**: large
- **EXCLUDED!**

Curating source corpora

Wikipedia

- **License**: CC BY-SA 3.0
- **Domain**: Wikipedia
- **Style**: formal, modern
- **Ethical risks**: low
- **Volume/Size**: large
- **INCLUDED!**

Curating source corpora

Airbnb Reviews

- **License**: CC0 1.0
- **Domain**: Review
- **Style**: colloquial, modern
- **Ethical risks**: medium
- **Volume/Size**: large
- **INCLUDED!**

Curating source corpora

Naver Entertainment News Reviews

- **License**: CC BY-SA 4.0
- **Domain**: Review
- **Style**: colloquial, modern
- **Ethical risks**: High
- **Volume/Size**: large
- **EXCLUDED!**

Curating source corpora

The Korean Economic Daily News

- **License**: CC BY-SA 4.0 for KLUE based on a **contract**
- **Domain**: News
- **Style**: Formal, modern
- **Ethical risks**: Low
- **Volume/Size**: large
- **INCLUDED!**

- **10 source corpora**
  - News Headlines
  - Wikipedia
  - Wikinews
  - Wikitree
  - Policy News
  - ParaKQC
  - Airbnb Reviews
  - NSMC
  - Acrofan News
  - The Korea Economics Daily News

Dataset	License	Domain	Style	Ethical Risks	Volume	Contemporary Korean
News Headlines	**N/A**	**News (Headline)**	**Formal**	**Low**	**Large**	**o**
Judgments	Public Domain	Law	Formal	Low	Large	o
National Assembly Minutes	Public Domain	Politics	Colloquial	Medium	Large	o
Patents	Public Domain	Patent	Formal	Low	Large	o
Wikipedia	**CC BY-SA 3.0**	**Wikipedia**	**Formal**	**Low**	**Large**	**o**
Wikibooks	CC BY-SA 3.0	Book	Formal	Low	Medium	x
Wikisource	CC BY-SA 3.0	Law Book	Formal	Low	Medium	x
Wikinews	**CC BY 2.5**	**News**	**Formal**	**Low**	**Small**	
Wikitree	**CC BY-SA 2.0**	**News**	**Formal**	**[illegible]**	[illegible]	[illegible]
Librewiki	CC BY-SA 3.0	Wiki	[illegible]	[illegible]	[illegible]	o
Zetawiki	CC BY-SA 3.0	[illegible]	[illegible]	[illegible]	Large	o
Policy News	**KOGL Ty[illegible]**	[illegible]	[illegible]	**Low**	**Medium**	**o**
NIKL Standard Korean Dicti[illegible]	[illegible]	Dictionary	Formal	Low	Large	o
[illegible]	CC BY-SA 2.0	Dictionary	Formal	Low	Large	o
ParaKQC	**CC BY-SA 4.0**	**Smart Home Utterances**	**Colloquial**	**Low**	**Medium**	**o**
Airbnb Reviews	**CC0 1.0**	**Review**	**Colloquial**	**Medium**	**Large**	**o**
NAVER Sentiment Movie Corpus (NSMC)	**CC0 1.0**	**Review**	**Colloquial**	**Medium**	**Large**	**o**
NAVER Entertainment News Reviews	CC BY-SA 4.0	Review	Colloquial	High	Large	o
Acrofan News	**CC BY-SA 4.0 for KLUE-MRC by Contract**	**News**	**Formal**	**Low**	**Large**	**o**
The Korea Economics Daily News	**CC BY-SA 4.0 for KLUE-MRC by Contract**	**News**	**Formal**	**Low**	**Large**	**o**

All openly usable, available and modifiable!!

Cleaning the source corpora

- **Noisy text**
  - Remove hash tags, html tags, incorrect unicode characters, empty parentheses and consecutive blanks.
  - Remove any sentences with more than 10 Chinese/Japanese characters.
  - Templated parts from news articles are removed: copyright marks, etc.
- Toxic content
- Person identifying information (PII)

Cleaning the source corpora

- Noisy text
- **Toxic content**
  - Automatic detection/removal of hate speech and gender bias
  - Not perfect, and manual detection/removal in the annotation time
- Person identifying information (PII)

Cleaning the source corpora

- Noisy text
- Toxic content
- **Person identifying information (PII)**
  - Regular expression based matching
    - email addresses, URL and @-references
  - Others are detected and removed manually in the annotation time.

Task-specific considerations

Every task is unique

- Task format
- Annotation
- Cleaning
- Evaluation metrics
- Artifacts (spurious correlation)
- and, more task-specific considerations

Let's look at a few sample tasks!

Topic classification

Source corpus: News headlines

- **Task format**
  - Input: a sequence of tokens (words, subwords, characters, etc.)
  - Output: a single category to which the input belongs
- Annotation
- Cleaning
- Evaluation metrics
- Annotation artifacts

Topic classification

Source corpus: News headlines

- Task format
- **Annotation**
  - We can’t rely on existing category tags
    - clickbait categories, undeniable categories
  - Three annotations per headline from 13 select crowdworkers based on pilot runs
    - Keep only headlines that have a majority category (final: 63,892 headlines)
- Cleaning
- Evaluation metrics
- Annotation artifacts

Topic classification

Source corpus: News headlines

- Task format
- Annotation
- **Cleaning**
    - Crowdworkers are asked to report problematic headlines
        - 650 headlines with PII’s, 194 with toxic content
        - 2,515 with no suitable categories
        - Total 2,953 headlines are excluded
- Evaluation metrics
- Annotation artifacts

Topic classification

Source corpus: News headlines

- Task format
- Annotation
- Cleaning
- **Evaluation metrics**
    - Macro F1 score: the average of the category-wise F1 scores.
- Annotation artifacts

Semantic textual similarity
Source corpora: AIRBNB, POLICY, PARAKQC

- **Task format**
  - Input: a sentence pair
  - Output: either **[0, 5]** or {0 (dissimilar), 1 (similar)}
- Instance sampling
- Annotation
- Cleaning & Annotation Artifact
- Evaluation metrics

Semantic textual similarity

Source corpora: AIRBNB, POLICY, PARAKQC

- Task format
- **Instance sampling**
  - Random sampling of a pair of sentences: almost always relevant sentences
  - PARAKQC: we use metadata (intent and topic) to sample sentence pairs
  - AIRBNB & POLICY: round-trip translation, ROUGE-based greedy matching, etc.
- Annotation
- Cleaning & Annotation Artifact
- Evaluation metrics

Semantic textual similarity

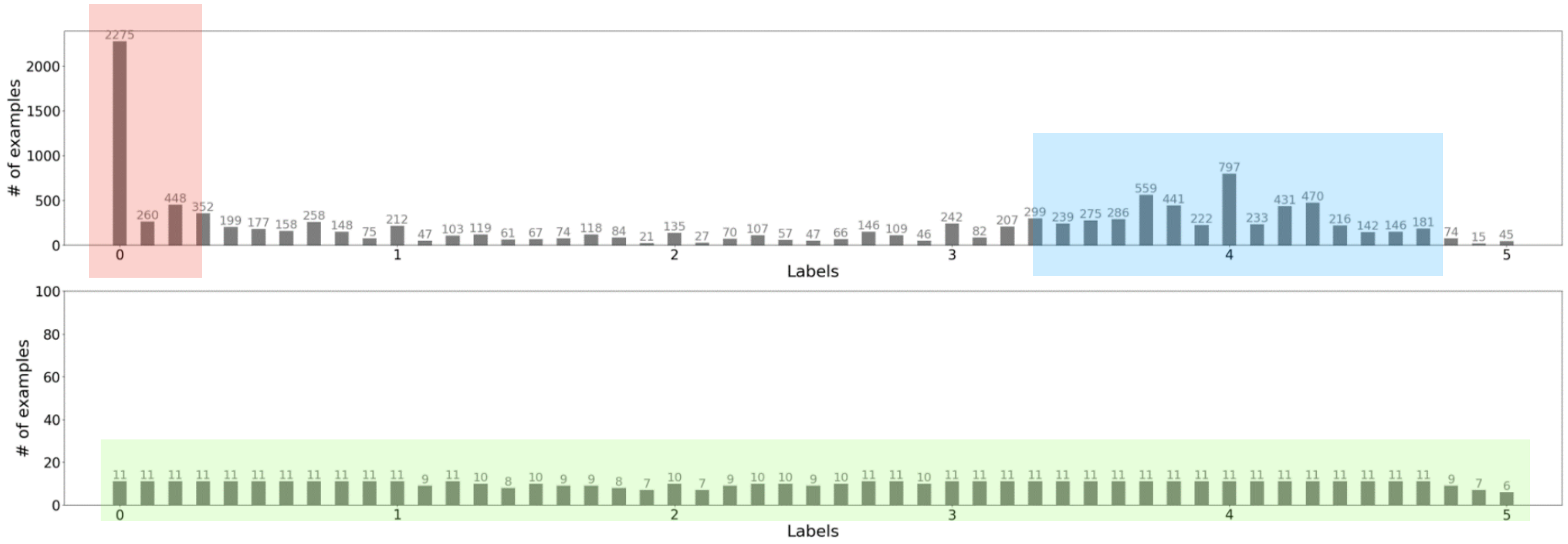
Source corpora: AIRBNB, POLICY, PARAKQC

- Task format
- Instance sampling
- **Annotation**
  - Started from SemEval-2015 but had to modify to fit Korean: [0, 5]
  - 19 select crowd workers for 14,869 sentence pairs
    - 2 annotators were excluded based on their score correlation against the other annotators.
  - at least 5 workers for each sentence pair: averaged and rounded to the first decimal point.
- Cleaning & Annotation Artifact
- Evaluation metrics

Semantic textual similarity

Source corpora: AIRBNB, POLICY, PARAKQC

- Task format
- Instance sampling
- Annotation
- **Cleaning & Annotation Artifact**



  - Crowd workers were asked to report any incorrect RTT: 418 pairs removed
  - Still skewed toward 0 and largely bimodal (peaks at 0 and 4)
  - Dev & test sets were constructed to be (largely) uniform over the score
- Evaluation metrics

Semantic textual similarity

Source corpora: AIRBNB, POLICY, PARAKQC

- Task format
- Instance sampling
- Annotation
- Cleaning & Annotation Artifact
- **Evaluation metrics**
    - Pearson’s correlation coefficient with continuous score
    - F1 score after binarizing the score (since the scores are largely bimodal)

Natural language inference

Source corpora: WIKITREE, POLICY, WIKINEWS, WIKIPEDIA, NSMC, AIRBNB

- **Task format**
  - Input: a sentence pair (premise, hypothesis)
  - Output: one of three categories {entailment, contradiction, neutral}
- Annotation
- Annotation Artifact

Natural language inference

Source corpora: WIKITREE, POLICY, WIKINEWS, WIKIPEDIA, NSMC, AIRBNB

- Task format
- **Annotation**
  - 546 workers from 2,604 workers after the pilot phase.
  - A premise is sampled from the source corpora.
  - A crowd worker writes a hypothesis.
  - Multiple crowd workers validate each premise-hypothesis pair.
  - Keep only pairs for which a majority consensus was made.
  - 30,998 final pairs
- Annotation Artifact

Natural language inference

Source corpora: WIKITREE, POLICY, WIKINEWS, WIKIPEDIA, NSMC, AIRBNB

- Task format
- **Annotation**
    - Careful annotation leads to higher quality data
- Annotation Artifact

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 $\neq$ Gold Label	26.00%	**0.00%**

Natural language inference

Source corpora: WIKITREE, POLICY, WIKINEWS, WIKIPEDIA, NSMC, AIRBNB

- Task format
- Annotation
- **Annotation Artifact**
  - A major issue: hypothesis-label correlation
  - Train a large classifier on the hypothesis-only input
  - Build dev/test tests to contain examples that cannot be well-predicted by the hypothesis-only input.

Relation extraction

Source corpora: WIKIPEDIA, WIKITREE, POLICY

- **Task format**
  - Input: a sentence with two entities marked.
  - Output: one of the 30 relation classes (inc. *no_relation*)
- Annotation
- Evaluation metrics

Relation Class
no_relation
org:dissolved *org:founded* *org:place_of_headquarters* *org:alternate_names* *org:member_of* *org:members* *org:political/religious_affiliation* *org:product* *org:founded_by* *org:top_members/employees* *org:number_of_employees/members*
per:date_of_birth *per:date_of_death* *per:place_of_birth* *per:place_of_death* *per:place_of_residence* *per:origin* *per:employee_of* *per:schools_attended* *per:alternate_names* *per:parents* *per:children* *per:siblings* *per:spouse* *per:other_family* *per:colleagues* *per:product* *per:religion* *per:title*

Relation extraction

Source corpora: WIKIPEDIA, WIKITREE, POLICY

- Task format
- **Annotation**
  - Each candidate sentence is automatically/manually inspected for hate spech
  - Automatically detect named entities from each sentence
    - Detect as many entities (more than 2) from each sentence
    - Manually clean up incorrect boundaries and incorrect entities
- Evaluation metrics

Relation extraction

Source corpora: WIKIPEDIA, WIKITREE, POLICY

- Task format
- **Annotation**
  - A major challenge: *no_relation* is way too dominant.
    - Pick a random pair of entities from a sentence, and they are unlikely to be directly related to each other.
    - Over-sample entity pairs that appear in KB and Wikipedia's infoboxes.
    - For dev/test sets, we do not over-sample but use uniform-sampling
  - Relation classes are annotated manually using crowdsourcing.
- Evaluation metrics

Relation extraction

Source corpora: WIKIPEDIA, WIKITREE, POLICY

- Task format
- Annotation
- **Evaluation metrics**
  - *no_relation* is dominant
    - We need to avoid incentivizing models that predict only *no_relation* well.
  - Micro F1 score on true relations (≠*no_relation*)
  - AUPRC (including *no_relation*)

Baselines matter

Pretraining

Facilitates rapid research

- Since 2018, it's become a standard approach to finetune a large-scale, pretrained language model for various natural language understanding tasks.
- A new benchmark suite must serve two purposes:
  - Provide a set of benchmark tasks based on which we can track progress
  - Provide a strong set of baselines on which progress can be made
- KLUE pretrains and releases large-scale language models.

Pretraining corpora

Separate from source corpora

- Pretraining corpora must be constructed differently from source corpora
    - As much information about the common language use must be retained
        - We do not (manually nor automatically) filter out hate speech, socially biased content, etc., because
            - to build a detector of these content, our model must be aware of them
            - it is not trivial to detect these from a large-scale corpus

Pretraining corpora

Separate from source corpora

- Pretraining corpora must be constructed differently from source corpora
  - As much information about the common language use must be retained
    - We do not filter out hate speech, socially biased content, etc.
    - We pseudonymize PII's.

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

Pretrained models

Separate from source corpora

- Because we cannot guarantee licenses behind various text often crawled off the internet, we do not release the pretraining corpora but only the pretrained models.
  - **MODU**: A collection of Korean corpora distributed by National Institute of Korean Languages
  - **CC-100-Kor**: the Korean portion of CC-100
  - **NAMUWIKI**: a Korean web-based encyclopedia
  - **NEWSCRAWL**
  - **PETITION**: a collection of public petitions posted to the Blue House

Pretrained models

Separate from source corpora

- **Base architectures**: BERT and RoBERTa
- **Tokenization**: morpheme-based byte-pair encoding
- **Comparisons**
  - **Multilingual models**: mBERT, XLM-R
  - **Korean-specific models**: KR-BERT, KoELECTRA

Pretrained models

Separate from source corpora

- KLUE does **not** rank models by the simple average of all the scores
- KLUE-RoBERTa$_{\text{LARGE}}$ generally works best across all the tasks.
- Multilingual models generally underperform language-specific ones.

Model	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	$F1^S$
mBERT$_{\text{BASE}}$	81.55	84.66	76.00	73.20	75.14	87.51	57.88	53.82	90.30	86.66	44.66	55.92	35.46	88.63
XLM-R$_{\text{BASE}}$	83.52	89.16	82.01	77.33	80.73	91.37	57.46	54.98	89.20	87.69	27.48	53.93	39.82	89.61
XLM-R$_{\text{LARGE}}$	86.06	92.97	85.86	85.93	81.81	92.49	58.39	61.15	92.71	88.70	35.99	66.77	41.20	89.80
KR-BERT$_{\text{BASE}}$	84.58	88.61	81.07	77.17	75.37	90.42	62.74	60.94	89.92	87.48	48.28	58.54	45.33	90.70
KoELECTRA$_{\text{BASE}}$	84.59	92.46	84.84	85.63	**86.82**	**92.79**	62.85	58.94	92.90	87.77	59.82	66.05	41.58	89.60
KLUE-BERT$_{\text{BASE}}$	85.49	90.85	82.84	81.63	84.77	91.28	66.44	66.17	92.14	87.77	62.32	68.51	48.99	91.86
KLUE-RoBERTa$_{\text{SMALL}}$	84.30	90.50	83.92	79.12	84.99	91.10	60.85	58.76	89.32	87.74	57.79	63.78	45.65	91.22
KLUE-RoBERTa$_{\text{BASE}}$	85.12	92.41	84.60	84.97	85.13	91.52	66.66	67.74	90.31	88.30	68.52	74.02	47.48	91.55
KLUE-RoBERTa$_{\text{LARGE}}$	**86.42**	**93.37**	**85.89**	**89.43**	85.79	91.77	**69.59**	**72.39**	**93.32**	**88.72**	**76.78**	**81.43**	**50.49**	**92.11**

Pretrained models

Separate from source corpora

- Morpheme-based subword tokenization generally works better than BPE
- This suggests the importance of customizing toward each target language

	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**

Pretrained models

Separate from source corpora

- Pseudonymization does not hurt the downstream accuracies
- This suggests we should put more effort in protecting privacy already at the pretraining stage without worrying about the downstream accuracies.

	YNAT	KLUE-STS		KLUE-NLI	KLUE-NER		KLUE-RE		KLUE-DP		KLUE-MRC		WoS	
Pretraining Corpus	F1	R^P	F1	ACC	$F1^E$	$F1^C$	$F1^{mic}$	AUC	UAS	LAS	EM	ROUGE	JGA	$F1^S$
Original	**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**
Pseudonymized	83.39	91.11	82.85	78.50	**84.99**	**91.22**	62.79	62.96	92.02	88.02	**62.88**	**67.58**	46.21	91.23

Summary

Considerations

Open access

- Benchmark corpora were carefully sourced to be released with CC BY SA.
    - Publicly accessible and distributable
    - Freely modifiable
- These properties maximize the accessibility and make KLUE future-proof

Considerations

Cleaning

- Benchmark corpora are carefully annotated and constructed to be free of
  - Hate speech
  - Various undesirable social biases
  - Personally identifiable information
- Pretraining corpora (not released) are filtered to be free of
  - Personally identifiable information, via pseudonymization

Considerations

Baselines

- Strong baselines are released publicly together with KLUE in order to
  - avoid meaningless effort in reproducing various not-so-strong baselines
  - facilitate further advances beyond the existing state of the art

Considerations
Leaderboard

- Leaderboard serves as an important way to broadcast the progress

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.

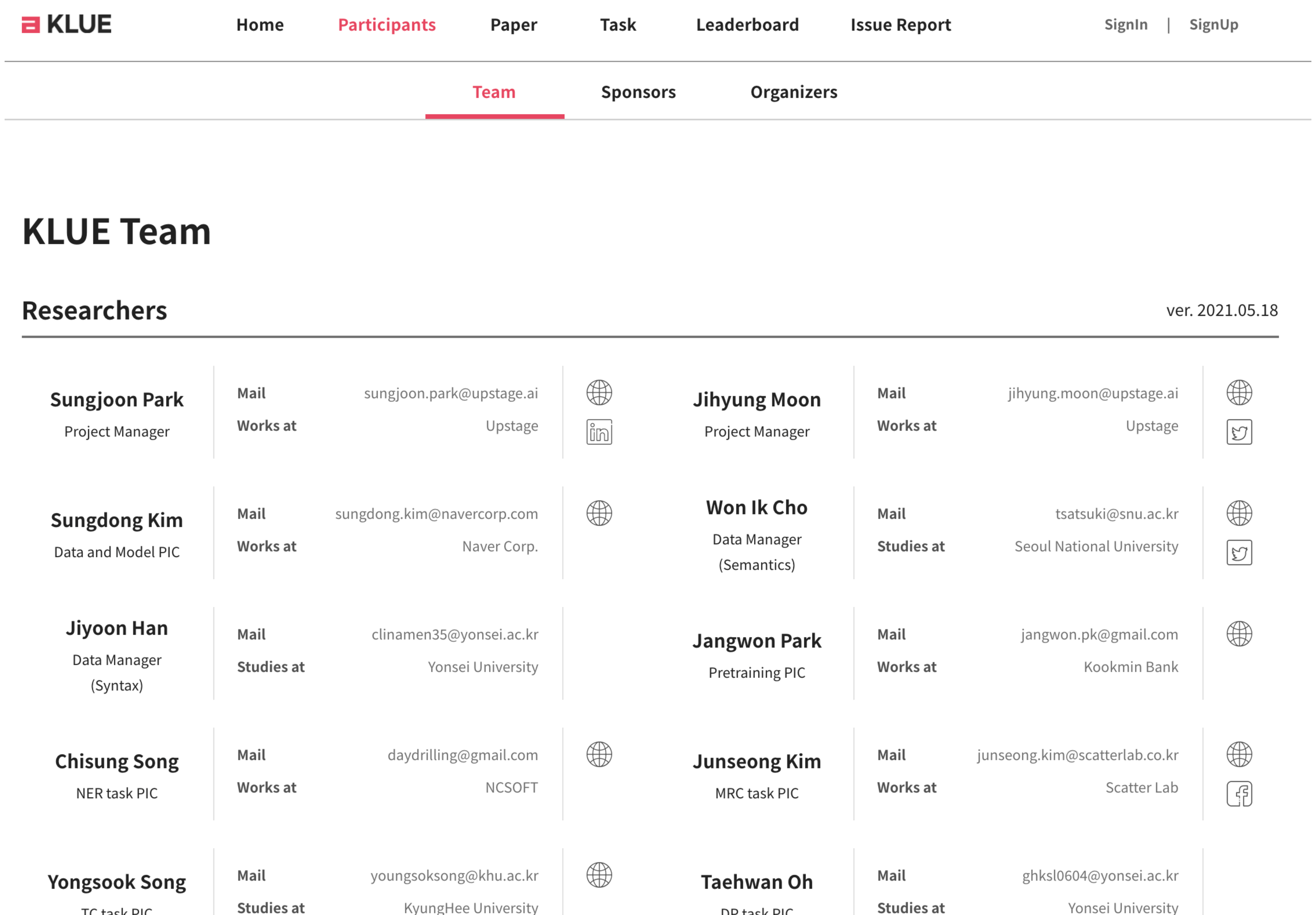
All | Small Size | Base Size | Large Size

#	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-RoBERTa-large	More	86.42	93.37	85.89	89.43	85.79	91.77	69.59	72.39	93.32	88.72	76.78	81.43	50.49	92.11
2	KLUE-team	KLUE-BERT-base	More	85.49	90.85	82.84	81.63	84.77	91.28	66.44	66.17	92.14	87.77	62.32	68.51	48.99	91.86
3	KLUE-team	KLUE-RoBERTa-base	More	85.12	92.41	84.6	84.97	85.13	91.52	66.66	67.74	90.31	88.3	68.52	74.02	47.48	91.55
4	KLUE-team	KLUE-RoBERTa-small	More	84.3	90.5	83.92	79.12	84.99	91.1	60.85	58.76	89.32	87.74	57.79	63.78	45.65	91.22

What it took to make KLUE

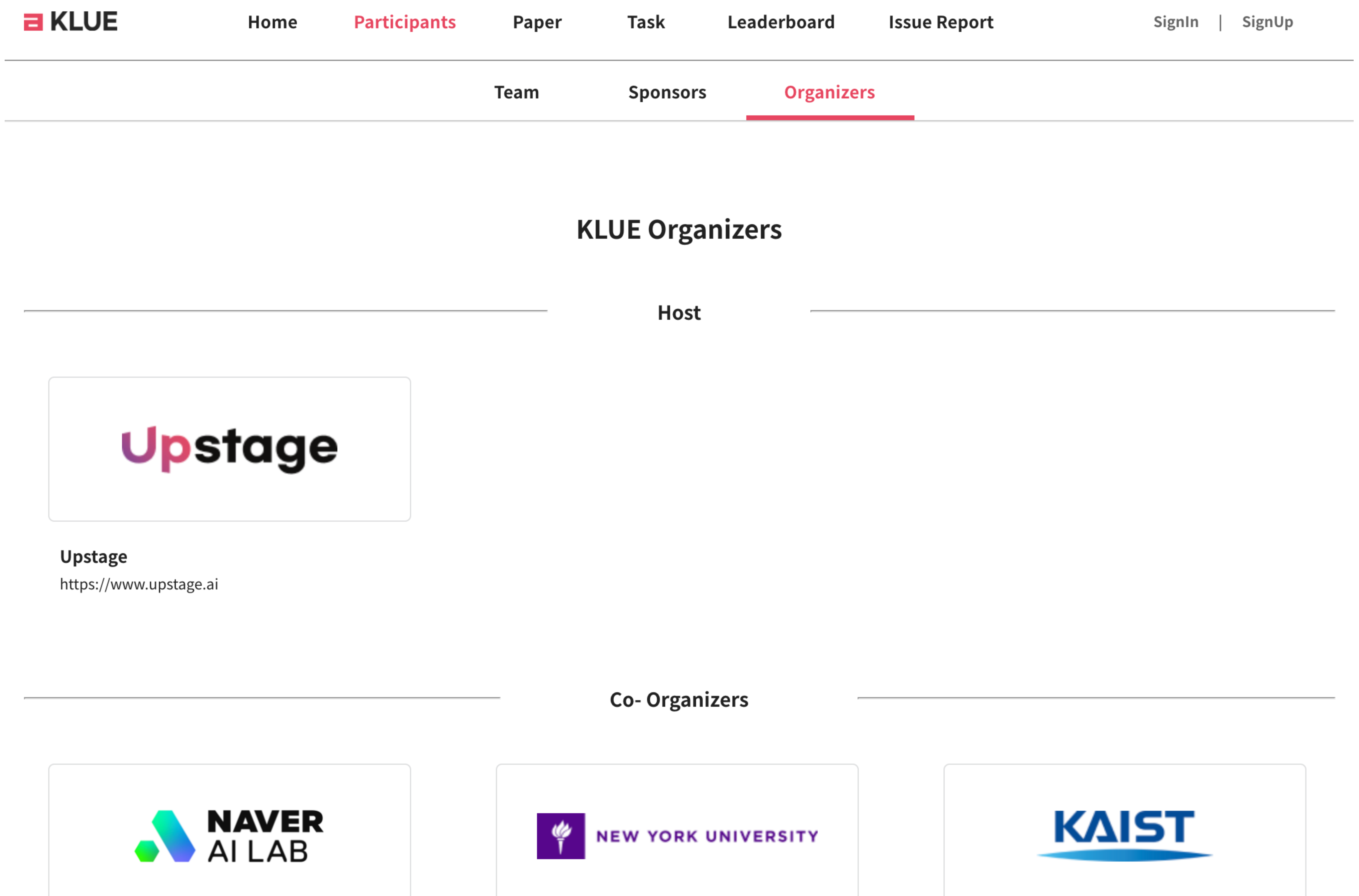
A collective effort

- 30+ researchers
  - NLP researchers
  - Crowdsourcing experts
  - ML researchers



A collective effort

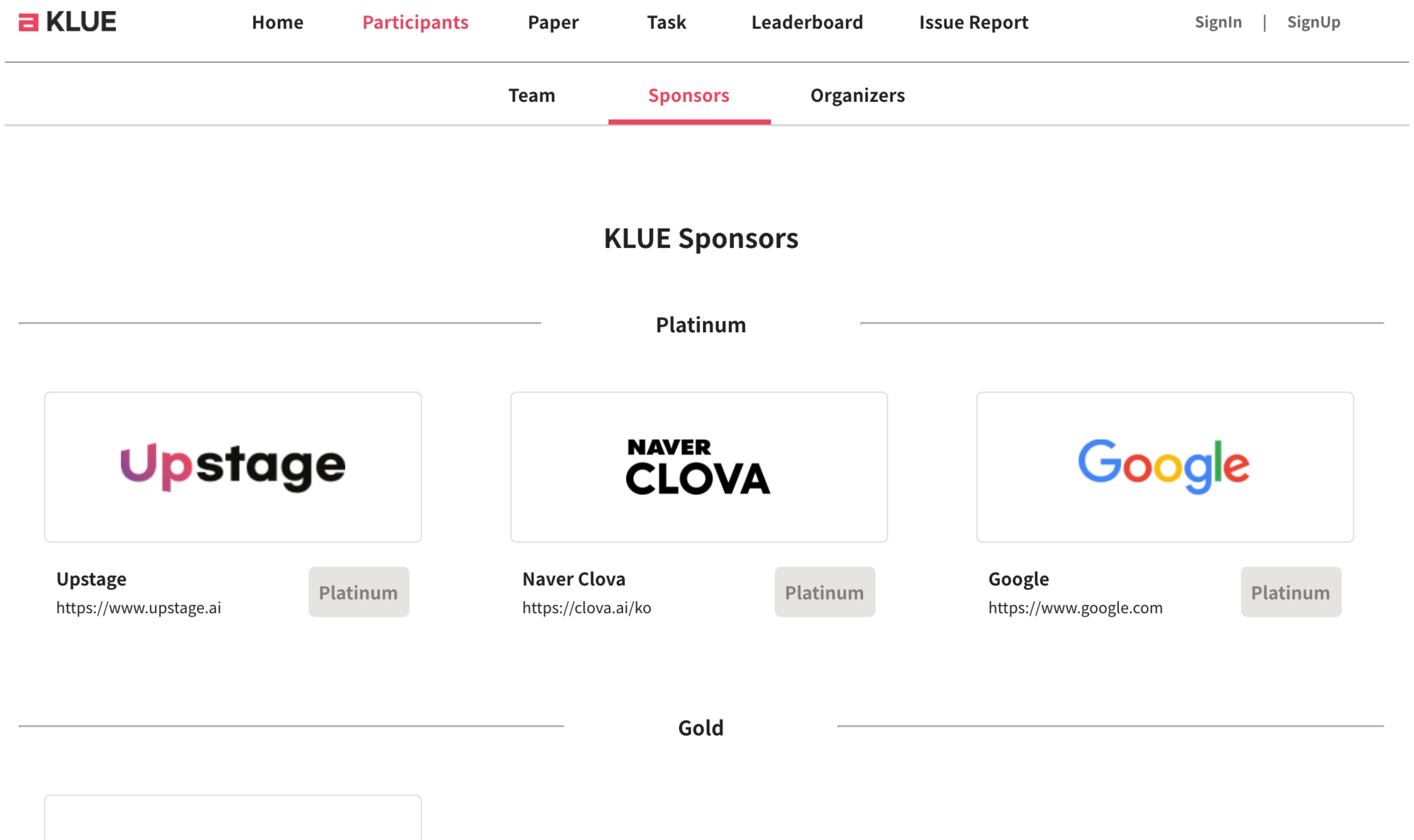
- From various organizations
  - Academic labs
  - Corporate labs
  - Crowdsourcing



Requires strong support

Industry and academia

- Researchers support
- Data support
- Compute support
- Annotation support
- Engineering support



We did it for Korean.

Let’s build one for your language!