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](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
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
- 

**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!**
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily News</td>
<td>10 source corpora</td>
<td>· The Korea Economics</td>
<td>· Acrofan News</td>
<td>· NSMC</td>
<td>· Airbn Reviews</td>
<td>· Policy News</td>
<td>· Wikipedias</td>
<td>· Wikinews</td>
<td>· News HEADLINES</td>
<td>· Wikipedia</td>
</tr>
<tr>
<td></td>
<td>· Daily News</td>
<td>10 source corpora</td>
<td>· The Korea Economics</td>
<td>· Acrofan News</td>
<td>· NSMC</td>
<td>· Airbn Reviews</td>
<td>· Policy News</td>
<td>· Wikipedias</td>
<td>· Wikinews</td>
<td>· News HEADLINES</td>
</tr>
<tr>
<td></td>
<td>· Daily News</td>
<td>10 source corpora</td>
<td>· The Korea Economics</td>
<td>· Acrofan News</td>
<td>· NSMC</td>
<td>· Airbn Reviews</td>
<td>· Policy News</td>
<td>· Wikipedias</td>
<td>· Wikinews</td>
<td>· News HEADLINES</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Statistics</th>
<th>KorNLI</th>
<th>KLUE-NLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unanimous Gold Label (4 Agree)</td>
<td>38.00%</td>
<td>71.00%</td>
</tr>
<tr>
<td>3 Agree with Gold Label</td>
<td>18.00%</td>
<td>24.00%</td>
</tr>
<tr>
<td>2 Agree with Gold Label</td>
<td>18.00%</td>
<td>3.00%</td>
</tr>
<tr>
<td>1 Agrees with Gold Label</td>
<td>16.00%</td>
<td>2.00%</td>
</tr>
<tr>
<td>0 Agrees with Gold Label</td>
<td>10.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Individual Label = Gold Label</td>
<td>64.50%</td>
<td>91.00%</td>
</tr>
<tr>
<td>No Gold Label (No 3 Labels Match)</td>
<td>4.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Majority Vote ≠ Gold Label</td>
<td>26.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>
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 extraction
Source corpora: WIKIPEDIA, WIKITREE, POLICY

• Task format

• Annotation
  • Each candidate sentence is automatically/manually inspected for hate speech
  • 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.

<table>
<thead>
<tr>
<th>Private Information</th>
<th>Pseudonymization</th>
<th>Pseudonymised Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telephone Number</td>
<td>Faker</td>
<td>055-604-8764</td>
</tr>
<tr>
<td>Social Security Number</td>
<td>Faker</td>
<td>600408-2764759</td>
</tr>
<tr>
<td>Foreign Registration Number</td>
<td>Faker</td>
<td>110527-1815659</td>
</tr>
<tr>
<td>Email Address</td>
<td>Faker</td>
<td><a href="mailto:agweon@example.org">agweon@example.org</a></td>
</tr>
<tr>
<td>IP Address</td>
<td>Faker</td>
<td>166.186.169.69</td>
</tr>
<tr>
<td>MAC Address</td>
<td>Faker</td>
<td>c5:d7:14:84:f8:cf</td>
</tr>
<tr>
<td>Mention(@)</td>
<td>Faker</td>
<td>@gildong</td>
</tr>
<tr>
<td>Address</td>
<td>Random Number Generation</td>
<td>경상북도 성남시 서초대64가</td>
</tr>
<tr>
<td>Bank Account Number</td>
<td>Random Number Generation</td>
<td>110-245-124678</td>
</tr>
<tr>
<td>Passport Number</td>
<td>Random Generation</td>
<td>M123A4567</td>
</tr>
<tr>
<td>Driver’s License</td>
<td>Random Number Generation</td>
<td>11-17-174133-01</td>
</tr>
<tr>
<td>Business Registration Number</td>
<td>Random Number Generation</td>
<td>123-45-67890</td>
</tr>
<tr>
<td>Health Insurance Information</td>
<td>Random Number Generation</td>
<td>1-2345678901</td>
</tr>
<tr>
<td>Credit or Debit Card Number</td>
<td>Random Number Generation</td>
<td>1234-5678-9012-3456</td>
</tr>
<tr>
<td>Vehicle Registration Place</td>
<td>Random Generation</td>
<td>55구 1601</td>
</tr>
<tr>
<td>Homepage URL</td>
<td>Random Generation</td>
<td><a href="http://www.example.com">www.example.com</a></td>
</tr>
</tbody>
</table>
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{\textsubscript{LARGE}} generally works best across all the tasks.
- Multilingual models generally underperform language-specific ones.

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

<table>
<thead>
<tr>
<th>Pretraining Corpus</th>
<th>YNAT F1</th>
<th>KLUE-STS R^P</th>
<th>KLUE-NLI F1</th>
<th>KLUE-NER F1^E</th>
<th>KLUE-RE F1^C</th>
<th>KLUE-DP F1^mic</th>
<th>KLUE-MRC UAS</th>
<th>KLUE-MRC LAS</th>
<th>WoS EM</th>
<th>WoS ROUGE</th>
<th>JGA</th>
<th>F1^S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>83.40</td>
<td>92.06</td>
<td>84.70</td>
<td>81.60</td>
<td>84.84</td>
<td>91.03</td>
<td>65.25</td>
<td>64.79</td>
<td>92.17</td>
<td>88.34</td>
<td>62.13</td>
<td>67.46</td>
</tr>
<tr>
<td>Pseudonymized</td>
<td>83.39</td>
<td>91.11</td>
<td>82.85</td>
<td>78.50</td>
<td>84.99</td>
<td>91.22</td>
<td>62.79</td>
<td>62.96</td>
<td>92.02</td>
<td>88.02</td>
<td>62.88</td>
<td>67.58</td>
</tr>
</tbody>
</table>

60
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.
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!