Cross-Level Semantic Similarity for Serbian Newswire Texts

Vuk Batanovič *, Maja Miličević Petrović †

* Innovation Center of the School of Electrical Engineering, University of Belgrade, Serbia
† Department of Interpreting and Translation, University of Bologna, Italy
Task: Cross-Level Semantic Similarity

- Input: two texts of different lengths, for example:
  - A sentence + a paragraph
  - A phrase + a sentence
- Output: a (fine-grained) semantic similarity score
- First formulated as a task in *SemEval 2014*
- So far, the only available annotated corpora for this task have been the *SemEval 2014* ones in English
A new annotated CLSS corpus in Serbian
  1000 sentence-paragraph + 1000 phrase-sentence pairs
Pairs were manually constructed using source texts gathered from a news aggregator website
  18,000 news articles written between June and August 2021
  Phrases (no finite verbs): news headlines
  Sentences (at least one finite verb): news subheads
  Paragraphs (at least two sentences): news article introductory paragraphs
CLSS.news sr annotation

- Similarity scores on a 0 - 4 Likert scale
- 5 annotators working in parallel
  - Each annotator constructed ~200 pairs per length pairing
    - Annotators instructed to aim for a balanced score distribution
  - Each annotator annotated all 1000 pairs per length pairing
  - Annotation guidelines - score definitions + 3 examples per score and length pairing
  - Final scores - average of individual annotators’ scores
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Lang.</th>
<th>Text pairs</th>
<th>Tokens</th>
<th>Avg. phrase length</th>
<th>Avg. sentence length</th>
<th>Avg. paragraph length</th>
<th>Vocab. size</th>
<th>Avg. similarity score</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>CLSS.news.nr</em> phrase-sentence</td>
<td>SR</td>
<td>1000</td>
<td>30K</td>
<td>~6</td>
<td>~23</td>
<td>/</td>
<td>12K</td>
<td>1.96</td>
</tr>
<tr>
<td><em>CLSS.news.nr</em> sentence-paragraph</td>
<td>SR</td>
<td>1000</td>
<td>86K</td>
<td>/</td>
<td>~22</td>
<td>~64</td>
<td>27K</td>
<td>1.91</td>
</tr>
<tr>
<td><em>SemEval CLSS</em> phrase-sentence</td>
<td>EN</td>
<td>1036</td>
<td>26K</td>
<td>~5</td>
<td>~19</td>
<td>/</td>
<td>8K</td>
<td>1.90</td>
</tr>
<tr>
<td><em>SemEval CLSS</em> phrase-sentence (newswire)</td>
<td>EN</td>
<td>425</td>
<td>13K</td>
<td>~5</td>
<td>~24</td>
<td>/</td>
<td>4K</td>
<td>1.76</td>
</tr>
<tr>
<td><em>SemEval CLSS</em> sentence-paragraph</td>
<td>EN</td>
<td>1034</td>
<td>93K</td>
<td>/</td>
<td>~19</td>
<td>~71</td>
<td>20K</td>
<td>1.84</td>
</tr>
<tr>
<td><em>SemEval CLSS</em> sentence-paragraph (newswire)</td>
<td>EN</td>
<td>301</td>
<td>26K</td>
<td>/</td>
<td>~20</td>
<td>~66</td>
<td>7K</td>
<td>1.68</td>
</tr>
</tbody>
</table>
Phrase-sentence pair distribution
Sentence-paragraph pair distribution
annotator self-agreement

- Measured on the initial ~200 pair sets
- Average self-agreement on:
  - Phrase-sentence pairs
    - Pearson $r = 0.930$
    - Spearman $\rho = 0.930$
    - Krippendorff’s $\alpha = 0.925$
  - Sentence-paragraph pairs
    - Pearson $r = 0.923$
    - Spearman $\rho = 0.925$
    - Krippendorff’s $\alpha = 0.921$
CLSS.news sr
inter-annotator agreement

- Average agreement between an annotator and the mean scores of other annotators on:
  - Phrase-sentence pairs
    - Pearson $r = 0.938$
    - Spearman $\rho = 0.938$
    - Krippendorff’s $\alpha = 0.929$
  - Sentence-paragraph pairs
    - Pearson $r = 0.937$
    - Spearman $\rho = 0.934$
    - Krippendorff’s $\alpha = 0.922$

- Global (non-binary) agreement on:
  - Phrase sentence pairs
    - Krippendorff’s $\alpha = 0.898$
  - Sentence-paragraph pairs
    - Krippendorff’s $\alpha = 0.897$
Analysis of pairs uniformly scored by all annotators

- Score 4 - Overlap in personal names/numbers, shared common lexical words
  - Items from the smaller element almost all present in the larger element
- Score 3 - Similar to 4, with partly different or omitted info
- Score 2 - Different personal names and shared/similar common vocabulary, or vice versa
- Score 1 - Some similarities in common vocabulary, synonyms more present than overlaps
- Score 0 - No shared words apart from the grammatical ones
Model evaluation on CLSS.news.sr

- 10-fold CV with sorted stratification
- Baseline - word overlap
- We consider two pre-trained language models
  - *Multilingual BERT* - pre-trained on 104 languages
  - *BERTić* - pre-trained on Serbian and closely related languages
- Settings we consider
  - Number of fine-tuning epochs: 1, 3, or 5
  - Adding training data: different CLSS.news.sr length pairings, or STS.news.sr (sentence pairs)
<table>
<thead>
<tr>
<th>Model</th>
<th>Additional training data</th>
<th>Epochs</th>
<th>Phrase - sentence similarity</th>
<th>Sentence - paragraph similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Pearson $r$</td>
<td>Spearman $\rho$</td>
</tr>
<tr>
<td>Word overlap</td>
<td>/</td>
<td>/</td>
<td>0.6361</td>
<td>0.6430</td>
</tr>
<tr>
<td></td>
<td>/</td>
<td>1</td>
<td>0.8756</td>
<td>0.8736</td>
</tr>
<tr>
<td></td>
<td>/</td>
<td>5</td>
<td>0.9010</td>
<td>0.8990</td>
</tr>
<tr>
<td>Multilingual BERT</td>
<td>CLSS.news.sr</td>
<td>1</td>
<td>0.8902</td>
<td>0.8893</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>0.9100</td>
<td>0.9060</td>
</tr>
<tr>
<td></td>
<td>/</td>
<td>1</td>
<td>0.8830</td>
<td>0.8851</td>
</tr>
<tr>
<td></td>
<td>/</td>
<td>5</td>
<td>0.9000</td>
<td>0.8974</td>
</tr>
<tr>
<td>BERTiće</td>
<td>CLSS.news.sr</td>
<td>1</td>
<td>0.9193</td>
<td>0.9239</td>
</tr>
<tr>
<td></td>
<td>/</td>
<td>5</td>
<td>0.9483</td>
<td>0.9439</td>
</tr>
<tr>
<td></td>
<td>STS.news.sr</td>
<td>1</td>
<td>0.9272</td>
<td>0.9277</td>
</tr>
<tr>
<td></td>
<td>/</td>
<td>5</td>
<td>0.9524</td>
<td>0.9486</td>
</tr>
<tr>
<td></td>
<td>STS.news.sr</td>
<td>1</td>
<td>0.9231</td>
<td>0.9236</td>
</tr>
<tr>
<td></td>
<td>/</td>
<td>5</td>
<td>0.9479</td>
<td>0.9442</td>
</tr>
</tbody>
</table>
Takeaways

- **CLSS.news.sr** - 1st CLSS dataset for a language other than English
- Excellent annotator agreements and balanced distribution across similarity scores
- Dataset and annotation guidelines are publicly available
- BERTić is the currently optimal model for CLSS in Serbian
- More epochs and the use of topically similar additional training data increases model performances
Thank you!
Questions?

Vuk Batanović
vuk.batanovic@ic.etf.bg.ac.rs

Maja Miličević Petrović
maja.milicevic2@unibo.it