13th International Conference on Language Resources and Evaluation
LREC 2022
Marseille, France

Cross-Level Semantic Similarity for Serbian Newswire Texts

Vuk Batanović *, 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

CLSS.news.sr

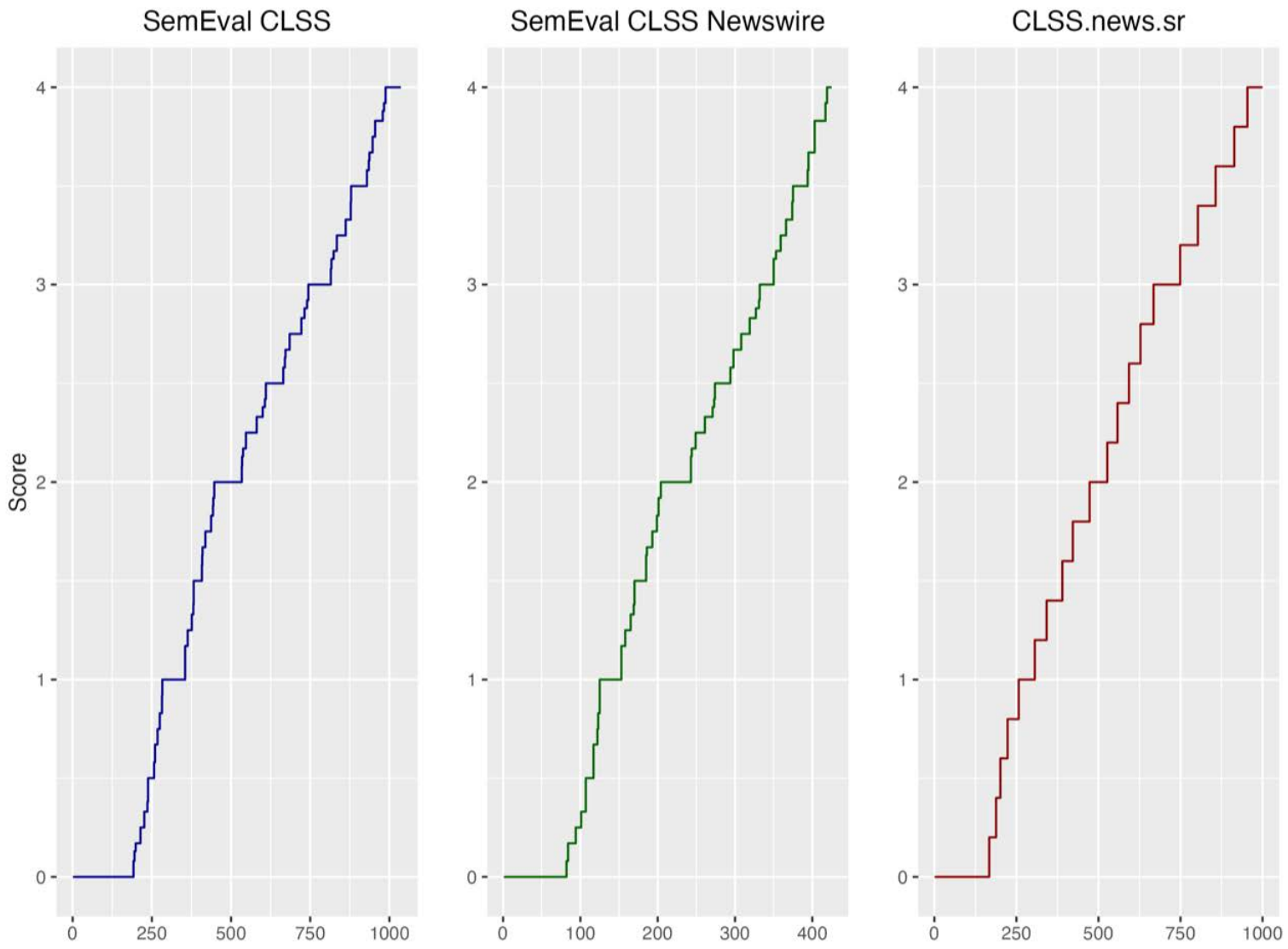
- A new annotated CLSS corpus in Serbian
  - 1000 sentence-paragraph + 1000 phrase-sentence pairs
- Pairs were manually constructed using source texts gathered from a news aggregator website
  - 18 000 news articles written between June and August 2021
  - Phrases (no finite verbs): news headlines
  - Sentences (at least one finite verb): news subheads
  - Paragraphs (at least two sentences): news article introductory paragraphs

CLSS.news.sr annotation

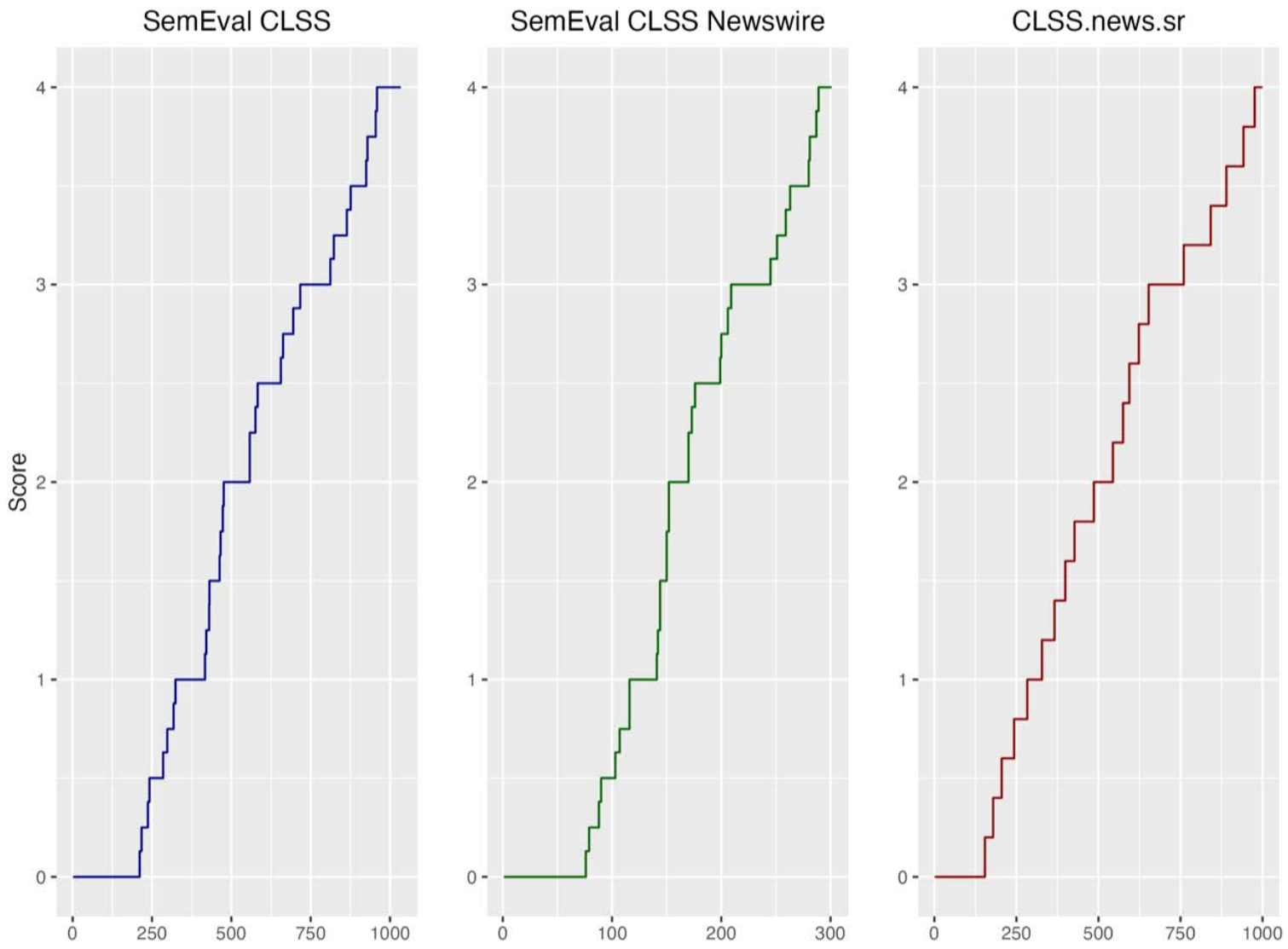
- Similarity scores on a 0 – 4 Likert scale
- 5 annotators working in parallel
  - Each annotator constructed ~200 pairs per length pairing
    - Annotators instructed to aim for a balanced score distribution
  - Each annotator annotated all 1000 pairs per length pairing
  - Annotation guidelines – score definitions + 3 examples per score and length pairing
  - Final scores – average of individual annotators' scores

Dataset	Lang.	Text pairs	Tokens	Avg. phrase length	Avg. sentence length	Avg. paragraph length	Vocab. size	Avg. similarity score
CLSS.news.sr phrase-sentence	SR	1000	30K	~6	~23	/	12K	1.96
CLSS.news.sr sentence-paragraph	SR	1000	86K	/	~22	~64	27K	1.91
SemEval CLSS phrase-sentence	EN	1036	26K	~5	~19	/	8K	1.90
SemEval CLSS phrase-sentence (newswire)	EN	425	13K	~5	~24	/	4K	1.76
SemEval CLSS sentence-paragraph	EN	1034	93K	/	~19	~71	20K	1.84
SemEval CLSS sentence-paragraph (newswire)	EN	301	26K	/	~20	~66	7K	1.68

Phrase-sentence pair distribution



Sentence-paragraph pair distribution



CLSS.news.sr 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 $\alpha = 0.897$

CLSS.news.sr preliminary qualitative linguistic analysis

- 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)

Model	Additional training data	Epochs	Phrase – sentence similarity		Sentence – paragraph similarity	
			Correlation coefficient			
			Pearson r	Spearman ρ	Pearson r	Spearman ρ
Word overlap	/	/	0.6361	0.6430	0.6458	0.6833
Multilingual BERT	/	1	0.8756	0.8736	0.9048	0.8941
		5	0.9010	0.8990	0.9265	0.9126
	CLSS.news.sr other length pairings	1	0.8902	0.8893	0.9187	0.9056
		5	0.9100	0.9060	0.9322	0.9198
	STS.news.sr	1	0.8830	0.8851	0.9110	0.9004
		5	0.9000	0.8974	0.9261	0.9132
BERTić	/	1	0.9193	0.9239	0.9077	0.9000
		5	0.9483	0.9439	0.9465	0.9334
	CLSS.news.sr other length pairings	1	0.9272	0.9277	0.9225	0.9135
		5	**0.9524**	**0.9486**	**0.9485**	**0.9368**
	STS.news.sr	1	0.9231	0.9236	0.9111	0.9008
		5	0.9479	0.9442	0.9405	0.9292

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
  - http://vukbatanovic.github.io/CLSS.news.sr/
- 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