# Slovene SuperGLUE Benchmark: Translation and Evaluation


## Aleš Žagar and Marko Robnik-Šikonja

University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, 1000 Ljubljana

ales.zagar@fri.uni-lj.si, marko.robnik@fri.uni-lj.si


## INTRODUCTION

- **The goal** of the SuperGLUE benchmark is to evaluate general language understanding.
- The tasks in SuperGLUE are diverse and comprised of **question answering** (QA), **natural language inference** (NLI), **coreference resolution** (CR), and **word sense disambiguation** (WSD).
- **Non-expert humans** evaluated all the tasks to provide a human performance baseline.
- It **significantly** contributes to the progress in the NLP area.
- However, many such benchmarks are available **only in English.**
- We **present and evaluate** a combined machine-human translation of SuperGLUE benchmarking suite for less-resourced Slovene language.

## DESCRIPTION OF TASKS

- **Boolean Questions** (BoolQ) is a QA task -> a question with a boolean answer related to a short text
- **CommitmentBank** (CB) is an NLI task-> 3-class agreement degree of a hypothesis and a short paragraph
- **The Choice Of Plausible Alternatives** (COPA) is a causal reasoning task-> one of the two sentences is the cause
- **Multi-Sentence Reading Comprehension** (MultiRC) is a QA task-> boolean answers to questions about text
- **Reading Comprehension with Commonsense Reasoning Dataset** (ReCoRD) is a multiple-choice QA task-> choose the correct entities from a paragraph to fill in a masked word in a given query
- **Recognizing Textual Entailment** (RTE) is an NLI task-> hypothesis entails a given text or not
- **Word-in-Context** (WiC) is a word sense disambiguation task-> sense matching of two polysemous words
- **Winograd Schema Challenge** (WSC) is a coreference resolution task-> two words refer or not to one entity

### Monolingual

| Task Models/Metrics | Avg | BoolQ Acc. | CB F1/Acc. | COPA Acc. | MultiRC $F1_a$/EM | ReCoRD F1/EM | RTE Acc. | WiC Acc. | WSC Acc. |
|---|---|---|---|---|---|---|---|---|---|
| Most Frequent | 45.7 | 62.3 | 21.7/48.4 | 50.0 | 61.1/0.3 | 33.4/32.5 | 50.3 | 50.0 | **65.1** |
| CBoW | 44.7 | 62.1 | 49.0/71.2 | 51.6 | 0.0/0.4 | 14.0/13.6 | 49.7 | 53.0 | **65.1** |
| BERT | 69.3 | 77.4 | 75.7/83.6 | 70.6 | 70.0/24.0 | **72.0/71.3** | 71.6 | **69.5** | 64.3 |
| BERT++ | 73.3 | **79.0** | **84.7/90.4** | **73.8** | **70.0/24.1** | **72.0/71.3** | **79.0** | **69.5** | 64.3 |
| Human (est.) | 89.8 | 89.0 | 95.8/98.9 | 100.0 | 81.8*/51.9* | 91.7/91.3 | 93.6 | 80.0 | 100.0 |
| Most Frequent | 49.1 | 63.3 | 21.7/48.4 | 50.0 | **76.4/0.6** | - | 58.6 | - | 65.8 |
| SloBERTa | 63.9 | 66.6 | **74.0/76.8** | **61.8** | 62.7/21.9 | - | **62.1** | - | **73.3** |
| CroSloEngual | 57.8 | 66.6 | 62.1/72.4 | 58.2 | 56.7/15.6 | - | **62.1** | - | 56.2 |
| mBERT | 59.1 | 70.0 | 66.6/73.6 | 54.2 | 57.4/16.3 | - | **62.1** | - | 61.6 |
| XML-R | 58.7 | **76.7** | 66.2/73.2 | 50.0 | 55.3/13.9 | - | 55.2 | - | 65.8 |

### Monolingual (HT vs. MT)

| Task Models/Metrics | Avg | BoolQ Acc. | CB F1/Acc. | COPA Acc. | MultiRC $F1_a$/EM | RTE Acc. |
|---|---|---|---|---|---|---|
| Most Frequent | 49.1 | 63.3 | 21.7/48.4 | 50.0 | **76.4/0.6** | **58.6** |
| HT-mBERT | 54.3 | 63.3 | 66.6/73.6 | 54.2 | 45.1/8.1 | 57.2 |
| MT-mBERT | 55.2 | 63.3 | 65.1/68.8 | 54.4 | 55.4/11.7 | 57.9 |
| HT-CroSloEngual | 55.6 | 63.3 | 62.1/72.4 | 58.2 | 53.0/8.4 | **58.6** |
| MT-CroSloEngual | 53.4 | 63.3 | 59.8/68.4 | 55.0 | 51.2/10.5 | 53.8 |
| HT-SloBERTa | **57.2** | 63.3 | **74.0/76.8** | **61.8** | 53.0/10.8 | 53.8 |
| MT-SloBERTa | 55.8 | 63.3 | 68.6/74.8 | 58.2 | 57.1/12.0 | 49.6 |
| HT-XLM-R | 53.5 | 63.3 | 66.2/73.2 | 50.0 | 53.3/0.9 | 57.2 |
| MT-XLM-R | 50.1 | 63.3 | 62.0/68.4 | 51.4 | 55.3/0.6 | 42.8 |
| HT-Avg | **55.1** | 63.3 | **70.6** | **56.0** | 29.1 | **56.7** |
| MT-Avg | 53.6 | 63.3 | 67.0 | 54.8 | **31.7** | 51.0 |

### Cross-lingual

| Evaluation | Model | source | target | Avg | BoolQ acc. | CB F1/acc. | COPA Acc. | MultiRC $F1_a$/EM | RTE Acc. | WSC Acc. |
|---|---|---|---|---|---|---|---|---|---|---|
| Zero-shot | CroSloEngual | english | slovene | 49.8 | 56.7 | 43.7/60.0 | 54.6 | 48.0/6.6 | 58.6 | 50.7 |
| | | slovene | english | 52.6 | 60.0 | 53.8/70 | **59.6** | **56.7/9.6** | 48.3 | 58.2 |
| | mBERT | english | slovene | 47.4 | 56.7 | 36.2/57.2 | 50.2 | 47.3/8.7 | 55.2 | 64.4 |
| | | slovene | english | 48.3 | 60.0 | 44.6/50.4 | 49.8 | 56.2/8.7 | 51.7 | 57.5 |
| | XLM-R | english | slovene | **53.8** | **63.3** | **62.9/68.4** | 53.6 | 48.5/0.3 | **62.1** | 56.2 |
| | | slovene | english | 51.7 | **63.3** | 59.1/67.2 | 47.2 | 52.9/12.9 | 51.7 | **65.8** |
| Few-shot | CroSloEngual | english | slovene | 54.4 | 60.0 | 52.4/68.6 | 55.0 | 52.8/9.72 | **65.5** | 54.1 |
| | | slovene | english | 53.0 | 60.0 | 53.8/70.0 | **59.5** | 56.0/12.1 | 49.7 | 58.2 |
| | mBERT | english | slovene | 50.9 | 60.1 | 53.1/66.2 | 50.4 | 50.8/9.8 | 53.8 | 64.4 |
| | | slovene | english | 51.3 | 60.7 | 51.8/58.2 | 50.3 | 57.2/11.1 | 56.5 | 56.8 |
| | XLM-R | english | slovene | **57.0** | **63.3** | **65.8/69.8** | 53.3 | **76.4/0.6** | 62.1 | 57.4 |
| | | slovene | english | 53.0 | **63.3** | 63.0/69.6 | 48.3 | 51.4/10.6 | 55.8 | **65.8** |
| | Most frequent | | | 52.4 | 63.3 | 23.0/52.7 | 50.0 | 77.3/0.3 | 58.6 | **65.8** |

### Main findings

# MAIN FINDINGS

- **SloBERTa** is the best model in monolingual evaluation setup.
- Performance **improves** with human translated datasets.
- **XLM-R** is the best model in cross-lingual scenarios.
- **All models** improved their average monolingual score in multilingual scenario.
- There is still a large room for **improvement** in all setups.

### Multilingual

| Evaluation on | Model | Avg | BoolQ Acc. | CB F1/acc. | COPA Acc | MultiRC F1a/EM | RTE Acc. | WSC Acc. |
|---|---|---|---|---|---|---|---|---|
| Slovene | CroSloEngual | 59.8 | 70.0 | 67.7/74.7 | 59.4 | 58.4/15.6 | 51.7 | 58.2 |
| | mBERT | 60.2 | 73.0 | 66.5/71.9 | 51.6 | 57.5/17.0 | 62.1 | 58.9 |
| | XML-R | 59.9 | 63.3 | 69.9/74.7 | 52.8 | 58.8/18.8 | 58.6 | 61.0 |
| English | CroSloEngual | 59.9 | 63.3 | 67.3/75.5 | **62.4** | 59.5/16.7 | 55.2 | 57.5 |
| | mBERT | **64.2** | **76.7** | 69.9/74.7 | 58.6 | **60.4/21.5** | **65.5** | 63.0 |
| | XML-R | 61.4 | 70.0 | **74.1/79.9** | 51.8 | 60.1/19.4 | 48.3 | **65.8** |
| | Most frequent | 52.4 | 63.3 | 23.0/52.7 | 50 | 77.3/0.3 | 58.6 | 65.8 |

## METHODS

### Translation procedure

- **Human** and **machine** translation-> approximately 120,000 words were human translated.
- Some datasets were **too large** (BoolQ, MultiRC, ReCoRD, RTE) and are partially human translated.
- We excluded ReCoRD and WiC, which need extensive manual editing (ReCoRD) or creation from scratch (WiC).

### Evaluated Slovene models

- **SloBERTa** -> Corpora training size: 3.47 billion tokens. Vocabulary size: 32k tokens. Pretraining task: whole word masking.
- **CroSloEngual** -> Corpora training size: 5.9 billion tokens (31% Croatian, 23% Slovenian, 47% English). Vocabulary size: 50k tokens. Pretraining task: whole word masking.

### Evaluation scenarios

- **Monolingual** -> train on Slovene, evaluate on Slovene
- **Cross-lingual** -> train on English, evaluate (in zero-shot and few-shot variants) on Slovene or English
- **Multilingual** -> train on both English and Slovene, evaluate on Slovene or English

## CONCLUSION

### Contributions

- We described the translation process and released Slovene version of SuperGLUE.
- We prepared a separate **Slovene leaderboard**, available at https://slobench.cjvt.si/.
- We encouraged the NLP community to pay attention to **less-resourced languages**.

### Further work

- Create a Slovene version of the **WiC** task from scratch.
- Manually adapt the **ReCoRD** task.
- **Increase samples** in partially translated datasets.

## ACKNOWLEDGEMENTS


The work was supported by the Slovenian Research Agency (ARRS) core research programme P6-0411 and the Ministry of Culture of Republic of Slovenia through the project Development of Slovene in Digital Environment (RSDO). This paper is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).We thank the authors of SuperGLUE benchmark for sharing test set answers for some of the tasks.