National AI Strategy and LT in Slovakia

Miroslav Zumrík (*Slovak Academy of Sciences*)
Mária Bieliková, Marián Šimko (*KInIT*)

11th LRB Meeting
25 November 2021

KInIT
Kempelen Institute
of Intelligent Technologies

General characteristics

The Slovak AI landscape has been fragmented

while past individual initiatives

- have tended to have a lesser lasting impact
- have faced the absence of an unifying force
- lacked governmental support

which contributed e.g. to the fact that Slovakia has not yet become member in international research infrastructures, such as CLARIN or DARIAH

- this remains a goal in the foreseeable future
- while it also changed in positive direction

Nevertheless, one can focus on academic and research centers throughout Slovakia

- universities in Košice, Žilina, Bratislava
- institutes within the *Slovak Academy of Sciences* (*Ľ. Štúr Institute of Linguistics, Institute of Informatics*)
- independent research institutes (*Kempelen Institute of Intelligent Technologies*)
- commercial stakeholders (*Xolution, Nettle*)
- units at public administration bodies responsible for data collection and analysis (*Ministry of Investments, Regional Development and Informatization, Ministry of Economy, Ministry of Interior or Ministry of Health*)

What is definitely needed and aimed for

- lasting governmental support
- effective harmonization of stakeholders efforts
- organizational impetus and endurance capable of causing the wished snowball effect
- institutional and organizational frame
- all with the goal of creating a functioning stakeholders network/organizational body on national level

Important political frame documents aiming at improving the Slovak AI efforts

- Action plan for the digital transformation of Slovakia for 2019–2022
  (tools for Slovak language analysis and processing are among actions in the Action plan)
- 2030 Digital Transformation Strategy for Slovakia
- Recovery and resilience plan for Slovakia
  (Component 9 Research and innovations, Component 17 Digital Slovakia)
- Road map for research infrastructures
  (SK VI Roadmap 2020–2030)

Recovery and resilience plan for Slovakia

- Creating and publishing language datasets and models for various use cases and properties, e.g. multilingual, generative, computing effective model
- Developing tools for Slovak text analysis and creating a central repository for Slovak language tools and datasets (prerequisite for becoming a member of CLARIN ERIC)
- Research and development in domain of disinformation detection

Good signs to be noticed

- Individual meetings of stakeholders begin to resonate in longer terms (such as the last year’s workshop on NLP and AI in Slovakia, more than 100 participants, 50% academy, 40% industry, 10% public)

proving that

- Stakeholders are aware of the need for coordination of efforts and willing to collaborate
- Slovak public bodies understand the importance of support and claim willingness to support, that is, the Ministry moved this action to the recovery and resilience plan for Slovakia

Committee for ethics and regulations in AI established

- January 2021, established by Ministry of Investments, Regional development and Informatization
- 21 members from academia, industry and public sectors
- Stance on ethics and regulation questions including Artificial Intelligence Act
- Chaired by Mária Bieliková

Notable stakeholder 1:
Slovak National Corpus, Ľ. Štúr Institute of Linguistics

- with a variety of nation-wide used resources and tools, corpora, word embedding research, various branches of linguistics research, - reaching from corpus linguistics, dialectology, terminology, sociolinguistics, discourse analysis, speech analysis, language acquisition research
- Newer outputs at the institute (by Radovan Garabík):
  - notably improved diacritics-reconstruction tool
  - improved morphological and lemmatization
  - corpus of freely accessible texts (in preparation)
  - new version of word embeddings for Hungarian
  - ARANEA web corpora
- Possible impact of resources and tools

Notable stakeholder 2:
Kempelen Institute of Intelligent Technologies (KInIT)

- to a great deal oriented on NLP related research
- operates within whole 5 different AI sections
- also offers doctoral studies, thus preparing new generations of AI researchers in collaboration with external mentors from abroad

SlovakBERT

- First public large-scale Slovak-only neural language model
- RoBERTa architecture, 124M parameters
- Published in September 2021
- Important **milestone** for Slovak NLP (no prior public model of similar kind)
- A result of industry-academia collaboration (with Gerulata Technologies)

- Data: ~20GB (4.6B words) of clean Slovak text, mainly web-crawled

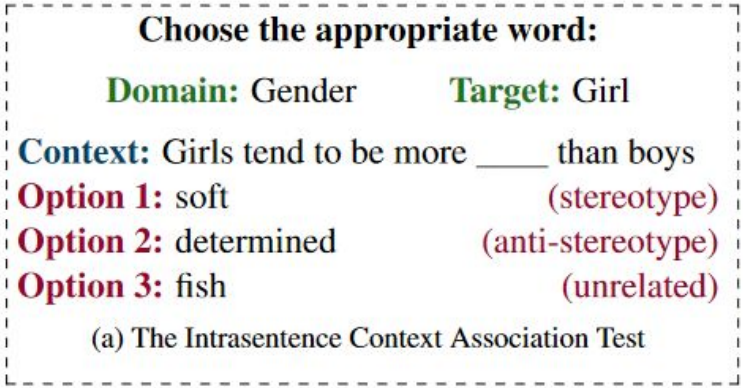
SlovakBERT

- Scientific evaluation (pre-print)
  - Tasks: POS tagging, Semantic similarity, Sentiment analysis, Document classification
  - State-of-the-art performance
  - Comparison with multilingual models (XLM-R-Large)
  - Model probing, model staleness
- Ready for and already involved in research projects with industry
  - APIs in development

SlovakBERT: Biases

Does the model contain unwanted societal stereotypes?

- Gathering stereotypes
  common for Slovak society
- Collaboration with human sciences
- Gender, ethnical groups, migrants, ...



(a) The Intrasentence Context Association Test

StereoSet

An impetus has been given – hopefully – in the right direction