Click here to be redirected to the virtual room of the Project Expo.

Project abbreviation: EMBEDDIA

Project name: Cross-Lingual Embeddings for Less-Represented Languages in European News Media

Project coordinator: Senja Pollak

Project consortium:

- Jožef Stefan Institute (SI) - Coordinator
- University of Ljubljana (SI)
- Queen Mary University of London (UK)
- University of Helsinki (FI)
- University of La Rochelle (FR)
- University of Edinburgh (UK)
- Trikoder – Styria Media Group (CRO)
- Ekspress Meedia (EE)
- Finnish News Agency STT (FI)
- TEXTA OÜ (EE)

Funding:

- ICT-29-2018: A Multilingual Next Generation Internet
- RIA: Domain-specific/challenge-oriented HLT
- 3m EUR

Project duration: 1. 1. 2019 - 31. 3. 2022

Main key words: cross-lingual embeddings, less-represented languages, deep neural networks, news media, comment filtering, news analysis, news generation

Background of the research topic: Access to the internet is a basic component of everyday life and civic engagement, but one in which language continues to be a challenge for fair and equitable access. As Europe becomes more multicultural, access to fundamental resources such as local news and government services is limited by the great diversity of the EU's 37 languages. Historically, the internet mostly developed in English, without clear awareness how language issues might form barriers to access and engagement, and without planning for multilingual support. In the EU, websites and online services for citizens offer resources in national local languages, and often only provide a second language (usually English) when absolutely needed. The great proliferation of web content, multiple and fast-changing content streams, and an expanding user interest make this approach untenable. Further, while advanced natural language processing (NLP) tools and resources exist for a few dominant languages (English, French, German), many of Europe's smaller language communities—and the news media industry that serves them—lack appropriate tools for multilingual internet development and multilingual news industry. We leverage recent advances in multilingual embeddings technologies that allow fast transfer of machine learning models from dominant to less-resourced languages.

Goal of the project: EMBEDDIA develops **multi-** and **cross-lingual embeddings technologies**, which allows cross-lingual analysis of texts and **fast transfer of machine learning models from dominant to less-resourced languages.** The main focus is on morphologically rich European languages (i.e. Croatian, Slovenian, Estonian, Finnish, Latvian, Lithuanian) and on applications for media industry including news analysis, comment analysis and news generation.

Click here to be redirected to the virtual room of the Project Expo.

Project abstract: For the EU to realise a truly equitable, open, multilingual online content and tools to support its management, new technologies allowing high quality transformations (not translations) between languages are urgently needed. The EMBEDDIA project (standing for Cross-Lingual Embeddings for Less-Represented Languages in European News Media) seeks to address these challenges by leveraging innovations in the use of **cross-lingual and multilingual embeddings** coupled with **deep neural networks** to allow existing monolingual resources to be used across languages, leveraging their high speed of operation for near real-time applications, without the need for large computational resources. Across three years (01/01/2019 to 31/12/2021), the project's six academic and four industry partners aim to develop novel solutions with focus on less-represented EU languages, and test them in **real-world news and media production** contexts.

Publications:

- Ulčar, Matej et al. (2021). **Evaluation of contextual embeddings on less-resourced languages**. *ArXiv preprint*: https://arxiv.org/abs/2107.10614
- Pollak, Senja et al. (2021): **EMBEDDIA Tools, Datasets and Challenges: Resources and Hackathon Contributions.** *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. https://aclanthology.org/2021.hackashop-1.14/
- Pelicon, Andraž et al. 2021. **Investigating cross-lingual training for offensive language detection**. *PeerJ computer science 7:e559*, doi: 10.7717/peerj-cs.559.
- Armendariz, Carlos et al. (2020). **SemEval-2020 task 3 : graded word similarity in context**. *Proceedings of the 14th International Workshop on Semantic Evaluation*, pages 36–49. https://aclanthology.org/2020.semeval-1.3.pdf
- Martinc, Matej et al. 2021. **TNT-KID : transformer-based neural tagger for keyword identification.** *Natural language engineering,* ISSN 1469-8110, doi: 10.1017/S1351324921000127.

For full list see http://embeddia.eu/outputs/