EMBEDDIA project: Cross-Lingual Embeddings for Less- Represented Languages in European News Media

Senja Pollak and Andraž Pelicon
Jozef Stefan Institute
Jamova 39, Ljubljana, Slovenia
senja.pollak,andraz.pelicon@ijs.si

Abstract

The EMBEDDIA project developed a range of resources and methods for less-resourced EU languages, focusing on applications for media industry, including keyword extraction, comment moderation and article generation.

1 Introduction

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. For the EU to realise a truly equitable, open, multilingual online content and tools to support its management, new multilingual technologies which do not rely on translation of text between languages are urgently needed. The aim of EMBEDDIA was 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 more than three years (01/01/2019 to 31/12/2021), *six academic partners* (Jozef Stefan Institute, the coordinating partner, Queen Mary University of London, University of Ljubljana, University of La Rochelle, University of Helsinki and University of Edinburgh) and *four industry partners* (TEXTA OÜ, As Ekspress Meedia, Finnish News Agency STT and Trikoder d.o.o.) developed novel solutions with focus on less-represented EU languages, and tested them in real-world media production contexts.

The main scientific goals of the project were to:

- Develop the *embeddings technology* for new generation NLP tools, which are both multilingual (able to deal with multiple languages) and cross-lingual (transfer easily across languages).
- Develop tools and resources for *less-resourced morphologically rich EU languages*, including Croatian, Estonian, Finnish, Latvian, Lithuanian and Slovenian.
- Leverage tools for well-resourced languages to be used for less-represented languages.

The project was strongly committed to address the challenges in news media industry, including:

- **Comment analysis** with mono- and cross-lingual applications in offensive speech filtering, fake news spreaders detection and sentiment analysis.
- **News analysis** with applications for keyword extraction, named entity recognition, news sentiment detection, viewpoints analysis, topic modelling, news linking, etc.
- **News generation** including text generation from structured data and headline generation.

1.1 Acknowledgements

This work has been supported by the EU's Horizon 2020 RIA under grant 825153 (EMBEDDIA), as well as ARRS core programme P2-0103.

2 Selected outputs

2.1 Datasets

EMBEDDIA has publicly released news and comments datasets (Pollak et al., 2021) in Estonian, Croatian, Russian and Latvian under the CC BY-NC-ND 4.0 license. We also created a set of novel benchmarks for evaluation, including CoSimLex dataset of word similarity in context (Armendariz et al., 2020) and cross-lingual analogy datasets (Ulčar et al., 2020).

2.2 Pretrained embeddings

Several monolingual and cross-lingual embeddings models have been trained for less-resourced EU languages (Ulčar et al., 2021), including ELMo embeddings, CroSloEngual BERT, LitLat BERT, FinEst BERT, SloRoberta and Est-Roberta.

2.3 Applications

Selected results include monolingual (Martinc et al., 2021), and cross-lingual (Koloski et al., 2022) keyword extraction methods, methods for cross-lingual offensive language detection (Pelicon et al., 2021), cross-lingual news sentiment analysis (Pelicon et al., 2020), cross-lingual Twitter sentiment detection (Robnik-Šikonja et al., 2020), named entity recognition (Boros et al., 2020), and article generation (Leppänen and Toivonen, 2021). Many other methods are described at `http://embeddia.eu/outputs/`.

3 Tools

The main EMBEDDIA tools are made available for future use through the EMBEDDIA Media Assistant, available at `https://embeddia.texta.ee/` consisting of:

- **API Wrapper**, intended for system integrations, including comment filtering, article analyzers and article generators.
- **Demonstrator**, showcasing a selection of the developed tools in a simple GUI for demonstration purposes (`https://embeddia-demo.texta.ee/`).
- **Tools Explorer** gathers a larger selection of tools relevant to media industry and research.
- **Texta Toolkit** GUI and API allow interactive user access and programming access to data exploration, investigative journalism and building own classifiers.

References

Armendariz, C. S., Purver, M., Ulčar, M., Pollak, S., Ljubešić, N., and Granroth-Wilding, M. (2020). CoSimLex: A resource for evaluating graded word similarity in context. In *Proc. of the 12th LREC*, pages 5878–5886, Marseille, France. ELRA.

Boros, E., Hamdi, A., Linhares Pontes, E., Cabrera-Diego, L. A., Moreno, J. G., Sidere, N., and Doucet, A. (2020). Alleviating digitization errors in named entity recognition for historical documents. In *Proc. of the 24th CoNLL*, pages 431–441, Online.

Koloski, B., Pollak, S., Škrlj, B., and Martinc, M. (2022). Out of thin air: Is zero-shot cross-lingual keyword detection better than unsupervised? In *arXiv:2202.06650*.

Leppänen, L. and Toivonen, H. (2021). A baseline document planning method for automated journalism. In *Proc. of the 23rd NoDaLiDa*, pages 101–111.

Martinc, M., Škrlj, B., and Pollak, S. (2021). Tntkid: Transformer-based neural tagger for keyword identification. *Natural Language Engineering*, page 1–40.

Pelicon, A., Pranjić, M., Miljković, D., Škrlj, B., and Pollak, S. (2020). Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, 10(17):5993.

Pelicon, A., Shekhar, R., Škrlj, B., Purver, M., and Pollak, S. (2021). Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7:e559.

Pollak, S., Robnik-Šikonja, M., Purver, M., Boggia, M., Shekhar, R., Pranjić, M., Salmela, S., Krustok, I., Paju, T., Linden, C.-G., et al. (2021). Embeddia tools, datasets and challenges: Resources and hackathon contributions. In *Proc. of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 99–109.

Robnik-Šikonja, M., Reba, K., and Mozetič, I. (2020). Cross-lingual transfer of Twitter sentiment models using a common vector space. In *Proc. of the Conference on Language Technologies & Digital Humanities*, pages 87–92.

Ulčar, M., Vaik, K., Lindström, J., Dailidėnaitė, M., and Robnik-Šikonja, M. (2020). Multilingual culture-independent word analogy datasets. In *Proc. of the 12th LREC*, pages 4074–4080, Marseille, France. ELRA.

Ulčar, M., Žagar, A., Armendariz, C. S., Repar, A., Pollak, S., Purver, M., and Robnik-Šikonja, M. (2021). Evaluation of contextual embeddings on less-resourced languages. *CoRR*, abs/2107.10614.