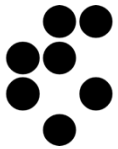
EMBEDDIA: Cross-Lingual Embeddings for Less-Represented Languages in European News Media

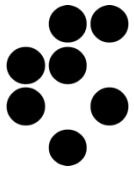
This project has received funding from European Union's Horizon 2020 research and innovation programme under grant agreement No 825153



Jožef Stefan Institute, Ljubljana, Slovenia

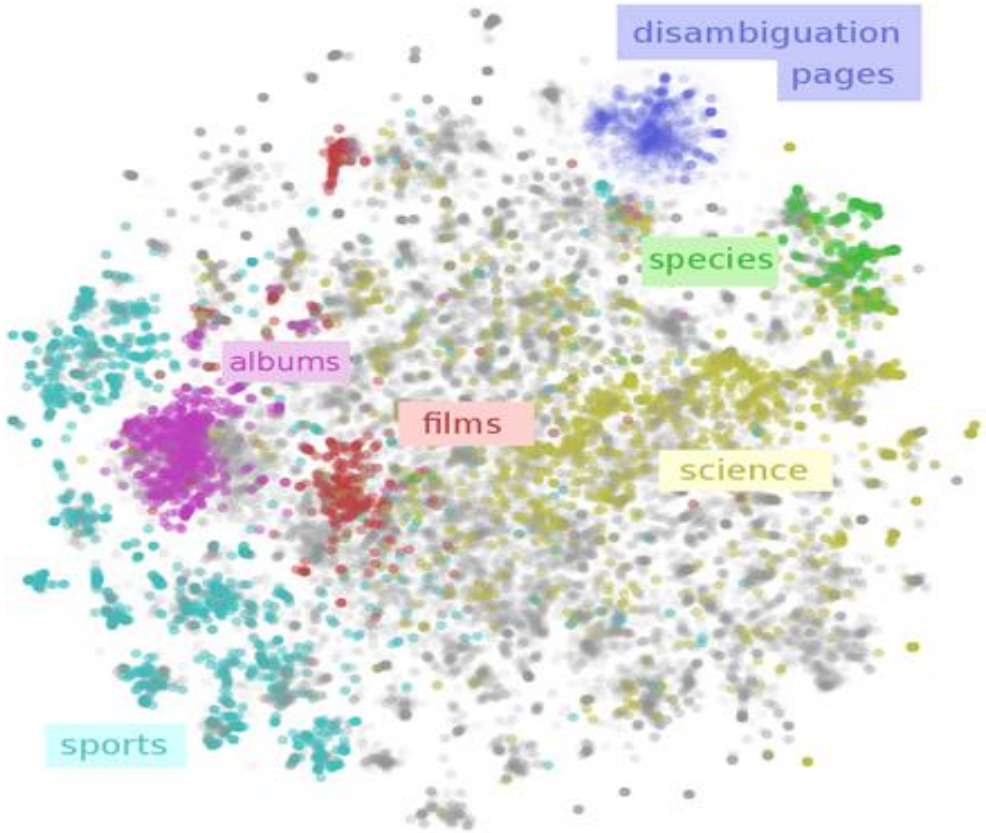
Project overview

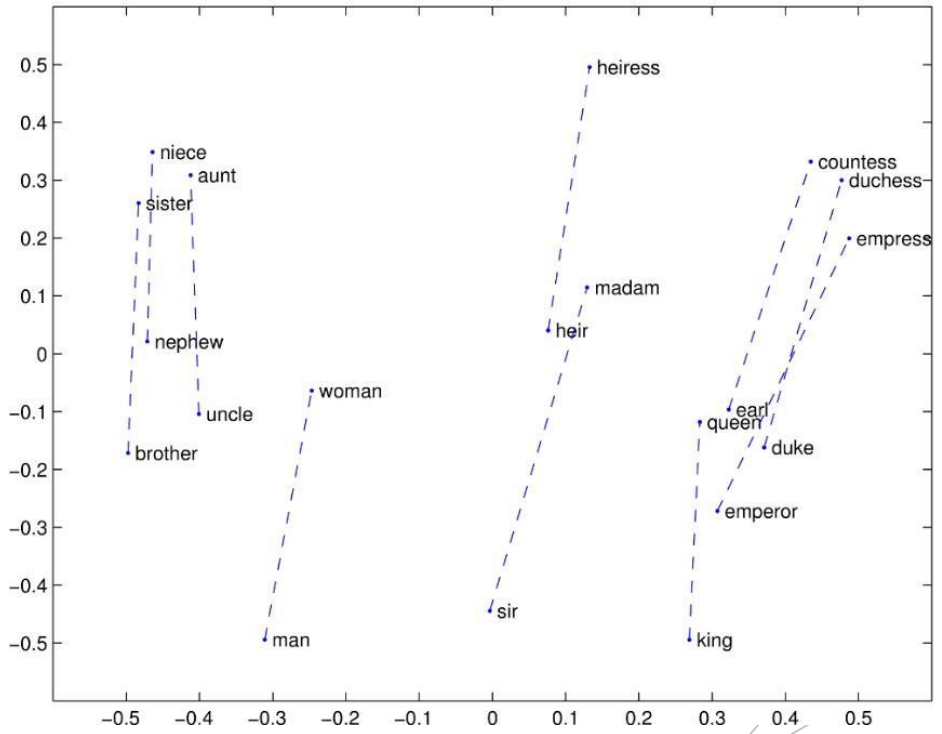
- **Cross-lingual embeddings** and deep neural networks enabling less-represented languages to benefit from resources and tools of well-resourced languages (English)
- Focus on morphologically-rich, **less-represented languages** in European news media: Estonian, Latvian, Lithuanian, Slovenian, Croatian and Finnish
- Applications for the **news media industry**: cross-lingual solutions for **news** and **user-generated content** analysis and **news generation**
    - Many tools that exist for English **do not exist or perform less well** for smaller languages



Word Embeddings

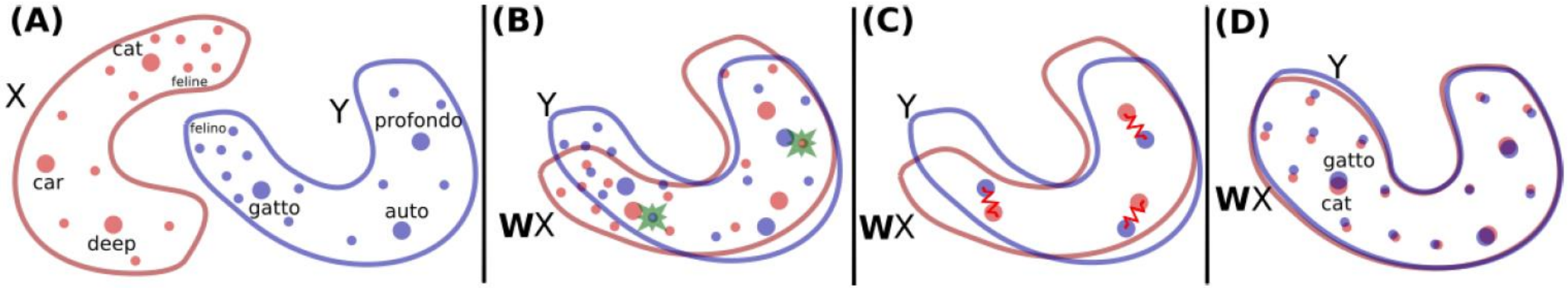
- Representations of word meaning from corpus statistics
- Spatial relationships correspond to linguistic relationships





Cross-Lingual Embeddings

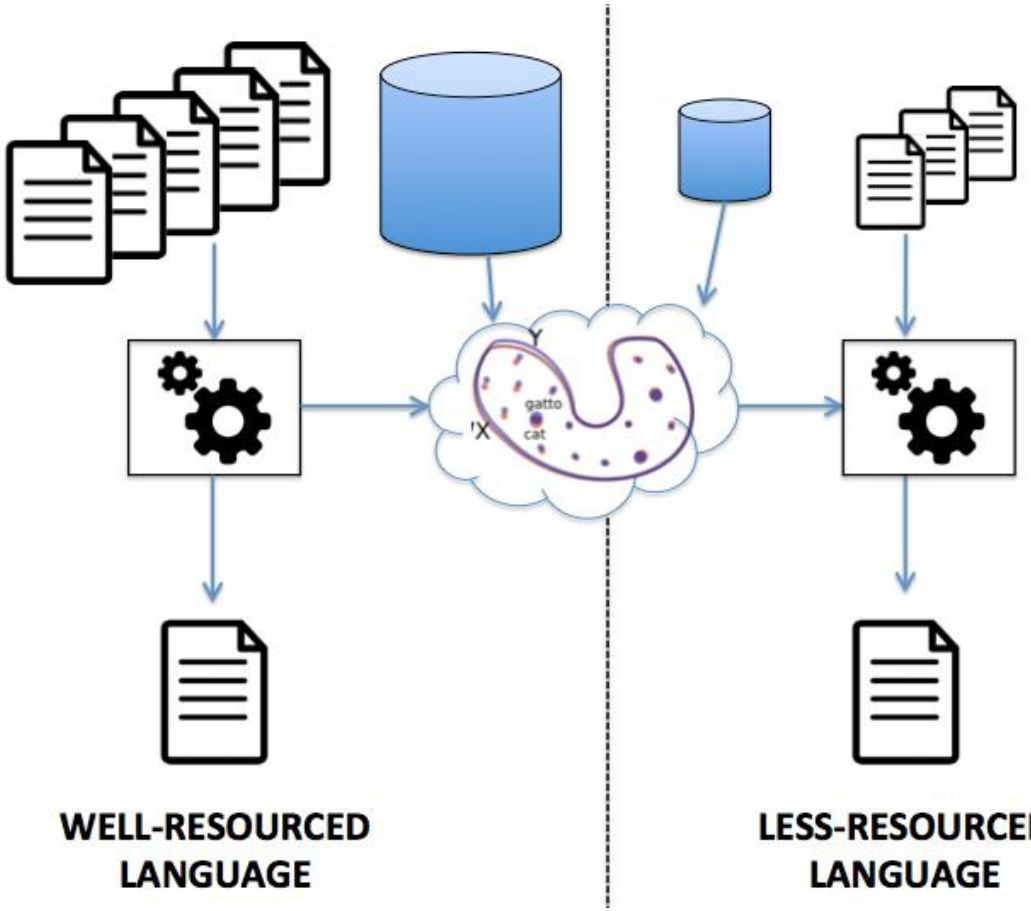
- Aligning embedding spaces across languages

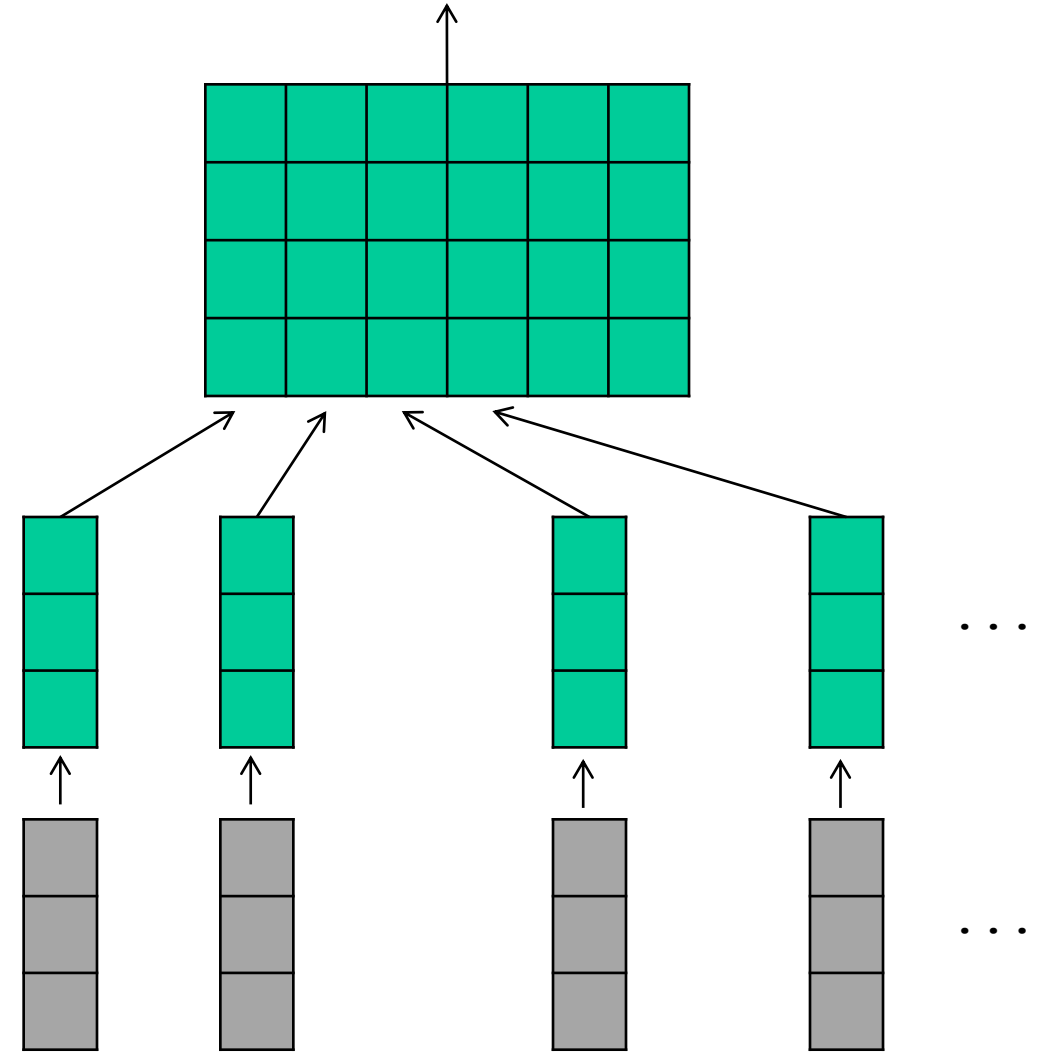


Conneau, A., Lample, G., Ranzato, M.A., Denoyer, L. and Jégou, H., 2018. Word translation without parallel data. Proceedings of ICLR 2018, also *arXiv preprint arXiv:1710.04087*.

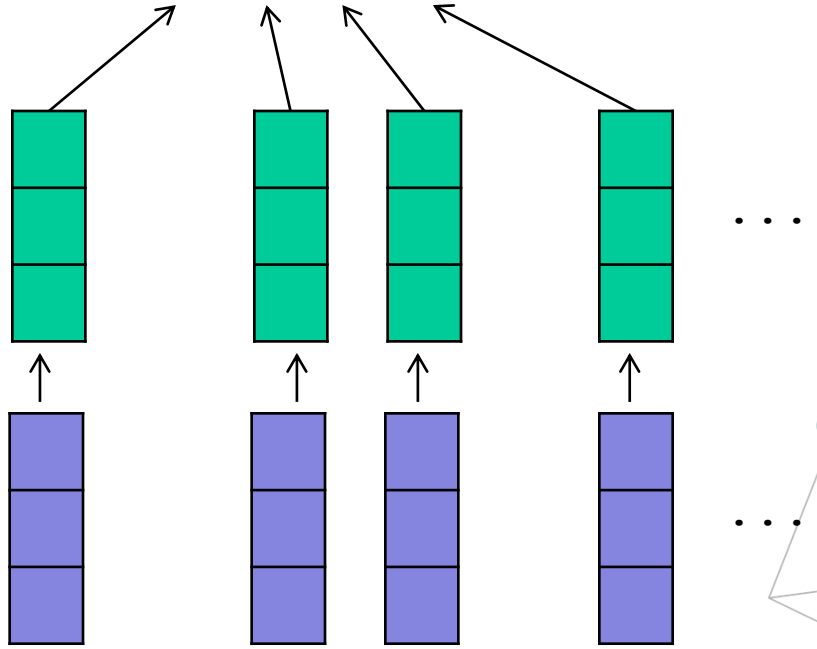
Cross-Lingual Embeddings

- Easy transfer of tools trained on mono-lingual resources





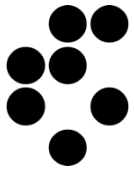
the president announced today ...



predsjednik je danas najavio ...

Solutions for Less-Resourced Languages

- EMBEDDIA technologies: **cross-lingual word embedding technologies**, coupled with **deep neural networks**, allow for rapid **transfer of tools across languages**
    - No need for large training sets, machine translation step, etc.
- Applications in challenges for the **news media industry**
    - Traditional news media companies around the world are confronting many **challenges** stemming from the **radical digital transformation of the media ecosystem**
    - New technologies influence each phase of media work from information sourcing to content packaging and distribution
- **EMBEDDIA will advance the technologies used in the news media industry for:**
    - **(i) cross-lingual news analysis;**
    - **(ii) automated multi-lingual news generation**;
    - **(iii) cross-lingual analysis of user-generated content (e.g. comments).**



Consortium

- **Interdisciplinary approach**: media studies, natural language processing and machine learning
- **Intersectoral consortium** (coordinated by Jožef Stefan Institute): 6 +2 academic partners, 3 news media partners, 1 SME

Academic partners:

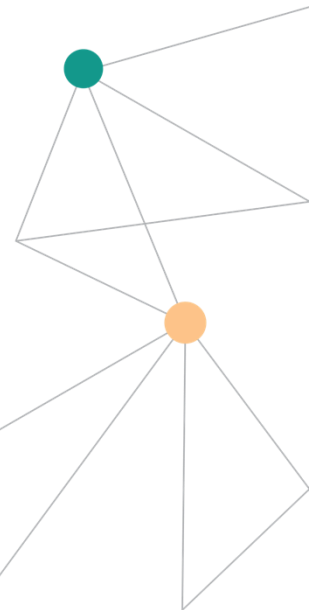
Jožef Stefan Institute (SI)
University of Ljubljana (SI)
Queen Mary Univ. of London (UK)
University of Helsinki (FI)
University of La Rochelle (FR)
University of Edinburgh (UK)

News media industry partners:

Trikoder (Styria) (HR)
Ekspress Meedia (EE)
Finnish News Agency STT (FI)

Text mining industry partner SME:

TEXTA OÜ (EE)



Project management team

Jožef Stefan Institute (SI)
Senja Pollak, project coordinator
Nada Lavrač, quality manager

University of Ljubljana (SI)
Marko Robnik-Šikonja, technical coordinator

Queen Mary University of London (UK)
Matthew Purver, data manager

External advisory board

Keaty Siivelt , advisor in the Government Office of Estonia, expert in Horizon 2020 “Secure Societies”

Hilde Van den Bulck, Professor of Communication Studies and Head of Department of Communication at Drexel University in Philadelphia

Ivan Vulic, Senior Research Associate, LT lab, the University of Cambridge.

Jef Verschueren, professor of linguistics, ex-dean and prof. Emeritus University of Antwerp

Language Technology Objectives

- Advance **cross-lingual word embedding technologies** (**WP1**)
- Advance corresponding basic **NLP technologies** (**WP2**)
- Advance cross-lingual **user comment analysis** (**WP3**): topic, stance, summarisation …
- Advance cross-lingual **news analysis** (**WP4**): linking, visualisation, viewpoint & bias …
- Advance multilingual **news generation** (**WP5**): content generation, creative language …
- Provide solutions via the **EMBEDDIA Media Assistant** (**WP6**)
- Provide **open access software (incl. ClowdFlows)** and exploit results (**WP7**)

- Example application: **detecting offensive comments**
  - Come and see our multilingual demo!