Multilingual Topic Labelling of News Topics using Ontological Mapping

Elaine Zosa, Lidia Pivovarova, Michele Boggia, and Sardana Ivanova
University of Helsinki
{firstname.lastname}@helsinki.fi

Training

IPTC news concepts (language-agnostic)

Articles (monolingual)

\[ W_1, W_2, \ldots, W_n \]

Multilingual SBERT encoder

\[ \text{i-hot encoding} \]

\[ X \]

MLP classifier

predicted concepts (associated with multilingual labels)

Testing/inference

Top topic terms

Top model (can be in multiple languages)

Sample outputs

**Topic (Finnish):** räikkönen, bottas, ajaa (*to drive*), hamilto, mercedes, ferrar

**Predicted labels:**
EN: motor racing/FI: autourheilu
EN: formula one/ FI: formula 1

**Topic (English):** film, movie star, director, hollywood, actor, minute

**Predicted labels:**
EN: film festival/FI: elokuvajuhlat
EN: cinema industry/FI: elokuvateollisuus

Objectives

- **Topic labelling** is the task of assigning concise labels to topics that captures its semantic content
- **Topic labels** aid topic interpretation and navigation of a corpus
- Multilingual news collections need to be more accessible to a broad range of users
- We need **multilingual** topic labels
- We need topic labelling methods that works on **topics in multiple languages**

Our method

- We cast topic labelling as a classification problem
- **Distant supervision:** top article words used as *pseudo topics* for training (Alokaili, et al. 2020)
- **Zero-shot inference:** produces labels for topics in languages not seen during training
- Produces **multilingual labels**
- Labels are on par with state-of-the-art label generation methods (see paper)

References


Acknowledgements

This work has been supported by the European Union’s Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (Embeddia).

More info!

Code & link to paper

https://github.com/ezosa/topic-labelling

References