multilingual_topic_labelling_of_news_topics_using_ontological_mapping

Year

2022

Author(s)

Zosa, Elaine and Pivovarova, Lidia and Boggia, Michele and Ivanova, Sardana

Title

Multilingual Topic Labelling of News Topics Using Ontological Mapping

Venue

ECIR

Topic labeling

Fully automated

Focus

Primary

Type of contribution

Novel

Underlying technique

Ontology-based (Ontological mapping method with SBERT embeddings)

Topic labeling parameters

Classification threshold (t): 0.03
Top document (training)/topic (inference) words provided as input (X): 30

Label generation

An ontological mapping method that maps topics to concepts in a language-agnostic news ontology and use the corresponding labels for these concepts (which are available in multiple languages) as topic labels.
Ontology mapping is approached as a multilabel classification task where a topic can be classified as belonging to one or more concepts in the ontology.

Following the distant-supervision approach in `alokaili_2020_automatic_generation_of_topic_labels` , we construct a dataset (see `Cropus` section) where the top n words of an article are treated as input X = (x1, . . . , xn) and the tagged concepts are the target C.
An article can be mapped to multiple concepts.
Top words can either be the top 30 scoring words by tf-idf (tfidf dataset) or the first 30 unique content words in the article (sent dataset). All models are trained on both datasets.

The classifier takes as an input a sequence X = (x1, . . . , xn) of the n top terms of a topic, and predicts P(c_i|X), the probabilities for each ontology concept c_i ∈ C.
The topic labels are obtained from the distribution P(c_i|X) as follows:

1. First, a list of label candidates is obtained by considering all c_i such that P(c_i|X) > t, where t is the classification threshold.
2. Then, we propagate the predicted concepts to the top of the ontology. For instance, if a topic is classified as belonging to concept 01005000:cinema, it also belongs to concept 01000000:arts, culture and entertainment, the parent of 01005000:cinema.
3. Lastly, we obtain the top topic labels by taking the most frequent concepts among the candidates and taking the labels of these concepts in the preferred language.

To compute the probabilities P(c_i|X), we encode the top terms (x1,...,xn) using SBERT and pass this representation to a classifier composed of two fully-connected layers with a ReLU non-linearity and a softmax activation.
We set the classification threshold t to 0.03 as determined by the validation set. We refer to this as the ontology model.

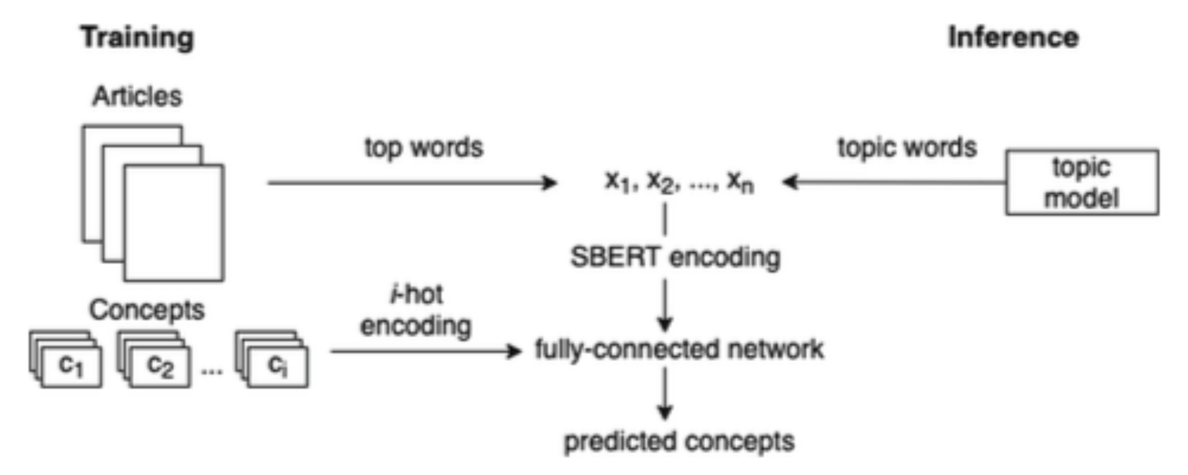

Fig. 1. News concepts prediction pipeline.

The paper also demonstrates the ability of the ontology models to label topics in a language it has not seen during training by testing it on English news topics.

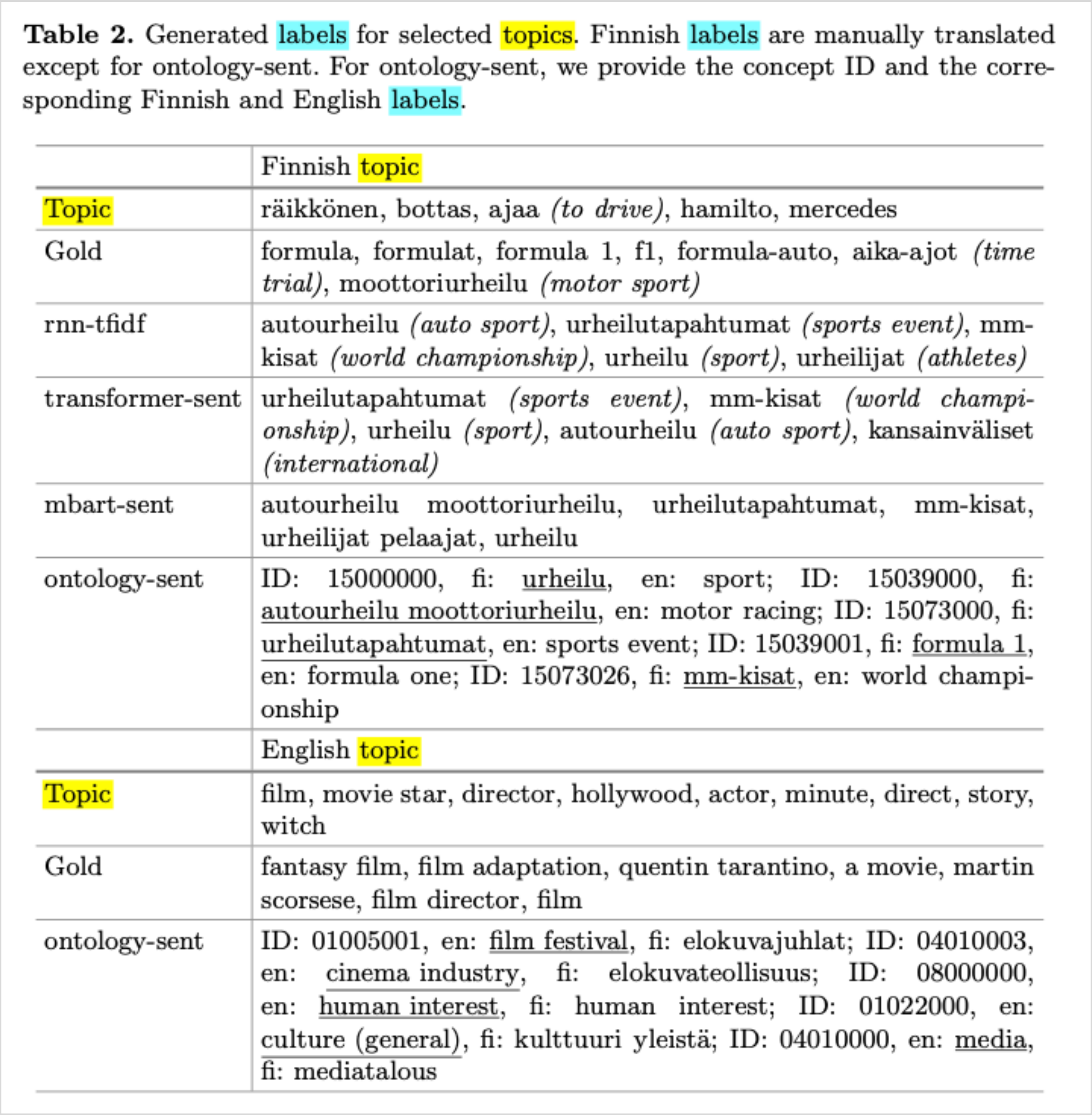

Table 2. Generated labels for selected topics. Finnish labels are manually translated except for ontology-sent. For ontology-sent, we provide the concept ID and the corresponding Finnish and English labels.

	Finnish topic
Topic	räikkönen, bottas, ajaa *(to drive)*, hamilto, mercedes
Gold	formula, formulat, formula 1, f1, formula-auto, aika-ajot *(time trial)*, moottoriurheilu *(motor sport)*
rnn-tfidf	autourheilu *(auto sport)*, urheilutapahtumat *(sports event)*, mm-kisat *(world championship)*, urheilu *(sport)*, urheilijat *(athletes)*
transformer-sent	urheilutapahtumat *(sports event)*, mm-kisat *(world championship)*, urheilu *(sport)*, autourheilu *(auto sport)*, kansainväliset *(international)*
mbart-sent	autourheilu moottoriurheilu, urheilutapahtumat, mm-kisat, urheilijat pelaajat, urheilu
ontology-sent	ID: 15000000, fi: urheilu, en: sport; ID: 15039000, fi: autourheilu moottoriurheilu, en: motor racing; ID: 15073000, fi: urheilutapahtumat, en: sports event; ID: 15039001, fi: formula 1, en: formula one; ID: 15073026, fi: mm-kisat, en: world championship
	English topic
Topic	film, movie star, director, hollywood, actor, minute, direct, story, witch
Gold	fantasy film, film adaptation, quentin tarantino, a movie, martin scorsese, film director, film
ontology-sent	ID: 01005001, en: film festival, fi: elokuvajuhlat; ID: 04010003, en: cinema industry, fi: elokuvateollisuus; ID: 08000000, en: human interest, fi: human interest; ID: 01022000, en: culture (general), fi: kulttuuri yleistä; ID: 04010000, en: media, fi: mediatalous

Motivation

Addressing the difficult and time-consuming task of manually interpreting labels by human annotators.

Topic modeling

LDA (applied on Test data 1)

Topic modeling parameters

Nr of topics (K): 100

Nr. of topics

30 topics generated from test data (top 30 from the original 100 with highest coherence)

Label

One or more concepts from the IPTC Subject Codes ontology

Label selection

\

Label quality evaluation

We use default topic labels–top five terms of each topic–as the baselines.

BERTScore (Zhang et al., 2019) is used to evaluate the labels generated by the models with regards to the gold standard labels.
BERTScore finds optimal correspondences between gold standard tokens and generated tokens
From these correspondences, recall, precision, and F-score are computed.
For each topic, the pairwise BERTScores between the gold labels and the labels generated by the models are computed and the maximum score is taken.
The scores are then averaged for all topics and report this as the model score.

Comparison baselines

- rnn:
  - RNN seq2seq model
  - 300-dim for the embedding layer and a hidden dimension of 200
  - Adam optimizer for 30 epochs with early stopping based on the validation loss
- transformer:
  - Slightly modified model where the RNN is replaced with transformers.
  - 6 layers for the encoder and decoder with 8 attention heads and an embedding dimension of 512
  - Adam optimizer for 30 epochs with early stopping based on the validation loss
- mbart:
  - Fine-tuned mBART-25 (Liu et al., 2020)

- Source and target languages set to Finnish
- mBART-25 from HuggingFace. Fine-tuned for 5 epochs. AdamW optimizer with weight decay set to 0.01.

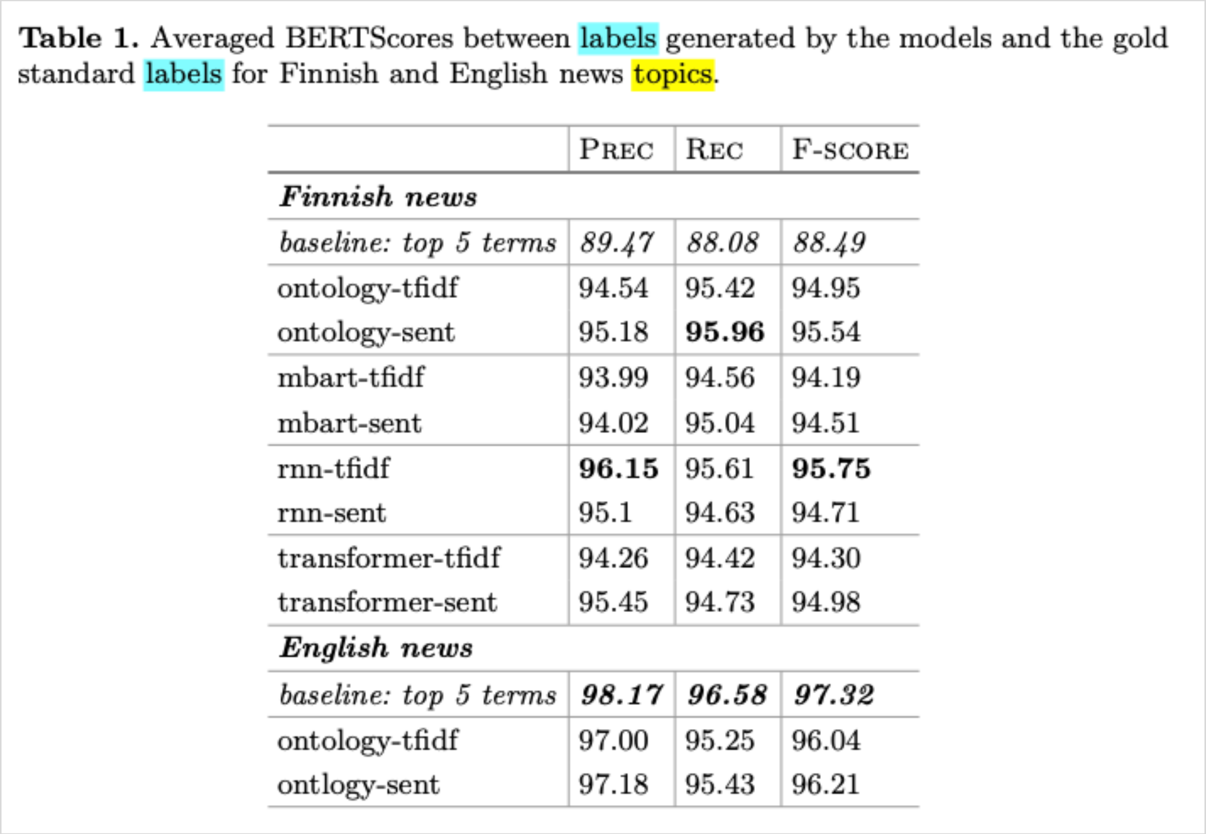

Table 1. Averaged BERTScores between labels generated by the models and the gold standard labels for Finnish and English news topics.

	PREC	REC	F-SCORE
Finnish news			
baseline: top 5 terms	*89.47*	*88.08*	*88.49*
ontology-tfidf	94.54	95.42	94.95
ontology-sent	95.18	**95.96**	95.54
mbart-tfidf	93.99	94.56	94.19
mbart-sent	94.02	95.04	94.51
rnn-tfidf	**96.15**	95.61	**95.75**
rnn-sent	95.1	94.63	94.71
transformer-tfidf	94.26	94.42	94.30
transformer-sent	95.45	94.73	94.98
English news			
baseline: top 5 terms	***98.17***	***96.58***	***97.32***
ontology-tfidf	97.00	95.25	96.04
ontlogy-sent	97.18	95.43	96.21

All models outperform the baseline by a large margin which shows that labels to ontology concepts are more aligned with human-preferred labels than the top topic words. The rnn-tfidf model obtained the best scores followed by ontology-sent. The transformer-sent and mbart-sent models also obtain comparable results. We do not see a significant difference in performance between training on the tfidf or sent datasets.

Assessors

\

Domain

Domain (paper): Topic labeling

Domain (dataset): News

Problem statement

Proposing a method to generate multilingual labels that capture the semantic content of a topic generated from a multilingual news collections.
Proposing an ontological mapping method that maps topics to concepts in a language-agnostic news ontology.
Proposing a method based on contextualised cross-lingual embeddings that works in a zero- shot setting, assigning labels to topics in languages not seen during training

Corpus

News ontology
Origin: IPTC Subject Codes Human readable concept data
Content: Three levels with 17 high-level concepts, 166 mid-level concepts and 1,221 fine-grained concepts. Mid-level concepts have exactly one parent and multiple children.
Details: This is a language-agnostic ontology designed to organise news content. Labels for concepts are available in multiple languages—in this work we focus specifically on Finnish and English.

Training data (Finnish)
Origin: Finnish News Agency (STT)
Nr. of documents: 385,803 article-concept pairs
Details:

- Finnish news
- 2017 STT articles (Finnish News Agency Archive 1 , Finnish News Agency Archive 2).
- Split 80/10/10 into train, validation and test sets.
- Two datasets are constructed by associating to each article (and concept) the top 30 words. Top words can either be:
    - The top 30 scoring words by tf-idf (tfidf dataset)
    - The first 30 unique content words in the article (sent dataset).

Test data 1 (Finnish)
Origin: Finnish News Agency (STT)
Content: 30 topics generated using LDA (see `Topic modeling` section)
Details:

- Finnish news
- 2018 STT articles

Test data 2 (English)
Origin: NETL dataset (Bhatia et al., 2016)
Content: 59 news topics with 19 associated labels

Details:

- English news topics with gold standard labels
- Used to test the model in a cross-lingual zero-shot setting, .
- Gold labels were obtained by generating candidate labels from Wikipedia titles and asking humans to evaluate the labels on a scale of 0-3.

Document

Training data (Finnish)

Finnish news article associated with ontological concepts and top 30 words (tfidf / sent)

Pre-processing

Training data (Finnish)

- Each article is tagged with IPTC concept(s)
- Each article is lemmatized with the Turku neural parser (Kanerva et al., 2013).

Following the distant-supervision approach in `alokaili_2020_automatic_generation_of_topic_labels`, we construct a dataset (see `Cropus` section) where the top n words of an article are treated as input X = (x1, . . . , xn) and the tagged concepts are the target C.
An article can be mapped to multiple concepts.
Top words can either be the top 30 scoring words by tf-idf (tfidf dataset) or the first 30 unique content words in the article (sent dataset). All models are trained on both datasets.

Test data 1 (Finnish)

- 30 topics generated using LDA (see `Topic modeling` section)
- **Gold standard labels generated topics generated from test data**
  - Three fluent Finnish speakers to provide labels for each of the selected topics.
  - For each topic, the annotators received the top 20 words and three articles closely associated with the topic.
  - The provided instructions are as follows: "Given the words associated with a topic, provide labels (in Finnish) for that topic. There are 30 topics in all. You can propose as many labels as you want, around 1 to 3 labels is a good number. We encourage concise labels (maybe 1-3 words) but the specificity of the labels is up to you. If you want to know more about a topic, we also provide some articles that are closely related to the topic. These articles are from 2018."
  - All unique labels are used as gold standard, which resulted in seven labels for each topic on average.

Test data 2 (English)

- Only the labels with a mean rating of at least 2.0 are taken as gold labels. Resulting in 330 topic-label pairs.

```
@inproceedings{zosa_2022_multilingual_topic_labelling_of_news_topics_using_ontological_mapping,
  abstract = {The large volume of news produced daily makes topic modelling useful for analysing topical trends. A topic is usually represented by a ranked list of words but this can be difficult and time-consuming for humans to interpret. Therefore, various methods have been proposed to generate labels that capture the semantic content of a topic. However, there has been no work so far on coming up with multilingual labels which can be useful for exploring multilingual news collections. We propose an ontological mapping method that maps topics to concepts in a language-agnostic news ontology. We test our method on Finnish and English topics and show that it performs on par with state-of-the-art label generation methods, is able to produce multilingual labels, and can be applied to topics from languages that have not been seen during training without any modifications.},
  address = {Cham},
  author = {Zosa, Elaine and Pivovarova, Lidia and Boggia, Michele and Ivanova, Sardana},
  booktitle = {Advances in Information Retrieval},
  date-added = {2023-03-03 14:31:17 +0100},
  date-modified = {2023-03-03 14:31:17 +0100},
  editor = {Hagen, Matthias and Verberne, Suzan and Macdonald, Craig and Seifert, Christin and Balog, Krisztian and N{\o}rv{\aa}g, Kjetil and Setty, Vinay},
  isbn = {978-3-030-99739-7},
  pages = {248--256},
  publisher = {Springer International Publishing},
  title = {Multilingual Topic Labelling of News Topics Using Ontological Mapping},
  year = {2022}}
```

#Thesis/Papers/Initial