Multilingual Topic Labelling of News Topics using Ontological Mapping

Zosa, Elaine

Springer
2022-04-05

Zosa , E , Pivovarova , L , Boggia , M & Ivanova , S 2022 , Multilingual Topic Labelling of News Topics using Ontological Mapping . in M Hagen , S Verberne , C Macdonald , C Seifert , K Balog , K Norvag & Setty (eds) , Advances in Information Retrieval. ECIR 2022 . þÿ L e c t u r e N o t e s i n C o m p u t e r S c i e n c e , v o l . 1 3 1 8 6 , S p r i n g e r , C h a m , p p Conference on Information Retrieval , Stavanger , Norway , 10/04/2022 . https://doi.org/10.1007/978-3-030-99739-7_

http://hdl.handle.net/10138/342489
https://doi.org/10.1007/978-3-030-99739-7_29

acceptedVersion

Multilingual Topic Labelling of News Topics using Ontological Mapping

Elaine Zosa[0000−0003−2482−0663], Lidia Pivovarova[0000−0002−0026−9902], Michele Boggia[0000−0002−4715−3691], and Sardana Ivanova[0000−0001−7819−435X]

University of Helsinki, Finland
firstname.lastname@helsinki.fi

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.

Keywords: topic labelling · ontology linking · cross-lingual embeddings

1 Introduction

Topic models uncover the latent themes in a document collection through the co-occurrences of words in documents [4]. The large volume of news produced daily makes topic models especially useful for tracking and analysing news trends [12, 14, 17]. A topic is usually represented by a ranked list of words but these words can be difficult and time-consuming to interpret for humans [10]. Therefore various methods have been proposed to assign concise labels to topics to improve interpretability [1, 3, 16, 18]. However, there has been no work so far on coming up with multilingual topic labels. Generating labels in multiple languages allows users to compare topical trends across linguistic boundaries without having to align topics and to explore news collections by users who might not have the necessary linguistic skills to do otherwise.

In this work we are interested in assigning concise multilingual labels to news topics. We propose an ontological mapping method that maps topics to concepts in a language-agnostic news ontology. These concepts have labels in multiple languages that we use as topic labels. We approach ontology mapping as a multilabel classification task where a topic can be classified as belonging to multiple concepts.

We train our classifier on a dataset of Finnish news and test it on Finnish and English topics, using the distant supervision approach proposed in Ref. [1], where articles are used as training data. Our method produces results that are on par with state-of-the-art label generation methods, produces multilingual labels and can be used for topics in languages that have not been used during training without any modification. The contributions in this paper are: (1) an ontological mapping approach that can produce topic labels in multiple languages; (2) 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; and (3) a novel dataset of Finnish news topics with gold standard labels.[1]

2 Related Work

Several existing methods for automatic topic labelling generate candidate labels either by extracting short phrases from topic-related documents [2, 9, 16] or from external sources such as Wikipedia [1, 9] and then ranking the candidates according to their relevance to the topic using distance metrics such as cosine distance [3] or the Kullback-Leibler divergence [8, 16].

Wikipedia is a popular external corpora for topic labelling, using article titles as candidate labels [3,9]. However, Ref. [9] argues that the broad domain covered by Wikipedia make it unsuitable for labelling topics from a domain-specific corpus, such as biomedical research papers. Moreover, Wikipedia sizes vary widely across different languages. Some previous work have also used ontologies [5, 7] but their methods rely on network analysis techniques to extract labels from the ontologies.

A more recent development is using deep learning to directly generate labels. Ref. [1] proposes a sequence-to-sequence model (seq2seq) trained on a synthetic dataset of Wikipedia articles and titles while Ref. [18] finetune BART, a pretrained transformer-based language model [11], with topic keywords and candidate labels from weak labellers to generate labels.

3 Experimental Setup

3.1 Models

Ontology Mapping. We propose an ontological mapping method that maps topics to concepts in a language-agnostic news ontology and use the corresponding labels for these concepts—available in multiple languages—as topic labels. We treat the ontology mapping problem as a multilabel classification task where a topic can be classified as belonging to one or more concepts in the ontology.

The classifier takes as an input a sequence $X = (x_1, \ldots, x_n)$ of the n top terms of a topic, and predicts $P(c_i|X)$, the probabilities for each ontology concept $c_i \in C$. The topic labels are obtained from the distribution $P(c_i|X)$ as follows: First, a list of label candidates is obtained by considering all c_i such that $P(c_i|X) > t$,

[1] Our code and dataset are available: https://github.com/ezosa/topic-labelling

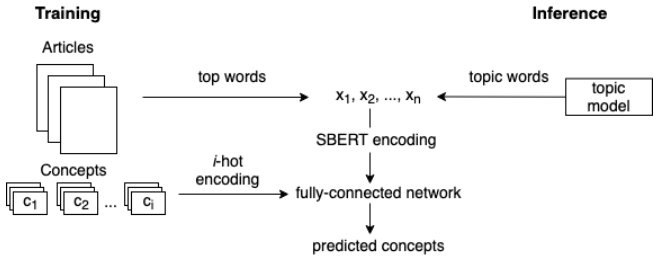

Fig. 1. News concepts prediction pipeline.

where t is the classification threshold. 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. 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 $(x_1, \ldots, x_n)$ using SBERT [19][2] 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. We illustrate this model in Figure 1.

Comparisons to State-of-the-art. We also investigate how our ontology mapping method compares to methods that directly generate topic labels. Ref. [1] uses an RNN-based encoder-decoder architecture with attention as a seq2seq model while Ref. [18] finetunes a pretrained BART model. Both methods have reported state-of-the-art results on English topics from multiple domains.

We implement a RNN seq2seq model using the same hyperparameters as [1]: 300-dim for the embedding layer and a hidden dimension of 200. We refer to this as the **rnn** model. We also implement a slightly modified model where we replace RNN with transformers, which has yielded state-of-the-art results in many NLP tasks. We use the hyperparameters from the original transformers model [22]: 6 layers for the encoder and decoder with 8 attention heads and an embedding dimension of 512. We refer to this as the **transformer** model.

Instead of BART which is trained only on English, we finetune a multilingual version, mBART [13], and set the source and target languages to Finnish. We finetuned mBART-25 from HuggingFace[3] for 5 epochs. We use the AdamW optimizer with weight decay set to 0.01. We refer to this as the **mbart** model [4]. For consistency, all the models except mbart are trained using Adam optimizer for 30 epochs with early stopping based on the validation loss.

[2] We use the multilingual model *distiluse-base-multilingual-cased*.
[3] https://huggingface.co/facebook/mbart-large-cc25
[4] While the mBART encoder is in a multilingual space, it cannot be used directly for cross-lingual language generation [15].

3.2 Datasets

News Ontology. We use the IPTC Subject Codes as our news ontology.[5] 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. This ontology has 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.

Training Data. We use news articles from 2017 of the Finnish News Agency (STT) dataset [20, 21] which have been tagged with IPTC concepts and lemmatized with the Turku neural parser [6]. Following the distant-supervision approach in [1], we construct a dataset where the top n words of an article are treated as input $X = (x_1, \ldots, x_n)$ 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. For each dataset, we have 385,803 article-concept pairs which we split 80/10/10 into train, validation and test sets.

Test Data. For Finnish topics, we train an LDA model for 100 topics on the articles from 2018 of the Finnish news dataset and select 30 topics with high topic coherence for evaluation. We also check that the topics are diverse enough such that they cover a broad range of subjects.

To obtain gold standard labels for these topics, we recruited 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. We provided the following instructions to the annotators:

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.

We reviewed the given labels to make sure the annotators understood the task and the labels are relevant to the topic. We use all unique labels as our gold standard, which resulted in seven labels for each topic on average. While previous studies on topic labelling mainly relied on having humans evaluate the labels outputted by their methods, we opted to have annotators *provide* labels instead because this will give us an insight into how someone would interpret a topic[6]. During inference, the input X are the top 30 words for each topic.

To test our model in a cross-lingual zero-shot setting, we use the English news topics and gold standard labels from the NETL dataset [3]. These gold labels were obtained by generating candidate labels from Wikipedia titles and asking humans to evaluate the labels on a scale of 0-3. This dataset has 59 news

[5] https://cv.iptc.org/newscodes/subjectcode/

[6] Volunteers are compensated for their efforts. We limited our test data to 30 topics due to budget constraints.

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

topics with 19 associated labels but we only take as gold labels those that have a mean rating of at least 2.0, giving us 330 topic-label pairs. We use default topic labels—top five terms of each topic—as the baselines.

4 Results and Discussion

We use BERTScore [23] 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 and from these correspondences, recall, precision, and F-score are computed. For each topic, we compute the pairwise BERTScores between the gold labels and the labels generated by the models and take the maximum score. We then average the scores for all topics and report this as the model score.

We show the BERTScores for the Finnish news topics at the top of Table 1. 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. In Table 2 (top), we show an example of the labels generated by the models and the gold standard labels. All models give sufficiently suitable labels, focusing on motor sports. However only the ontology-sent model was able to output 'formula 1' as one of its labels.

We also demonstrate 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 from the NETL dataset [3]. This dataset was also used in Ref. [1] for testing but our results are not comparable since they present the scores for topics from all domains while we only use the news topics. The results are shown at the bottom

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

of Table 1. Although the ontology models do not outperform the baseline, they are still able to generate English labels that are very close to the gold labels considering that the models have been trained only on Finnish data. From the example in Table 2 (bottom), we also observe that the gold labels are overly specific, suggesting names of directors as labels when the topic is about the film industry in general. We believe this is due to the procedure used to obtain the gold labels, where the annotators were asked to *rate* labels rather than propose their own.

5 Conclusion

We propose a straightforward ontology mapping method for producing multilingual labels for news topics. We cast ontology mapping as a multilabel classification task, represent topics as contextualised cross-lingual embeddings with SBERT and classify them into concepts from a language-agnostic news ontology where concepts have labels in multiple languages. Our method performs on par with state-of-the-art topic label generation methods, produces multilingual labels, and works on multiple languages without additional training. We also show that labels of ontology concepts correlate highly with labels preferred by humans. In future, we plan to adapt this model for historical news articles and also test it on more languages.

Acknowledgements

We would like to thank our annotators: Valter Uotila, Sai Li, and Emma Vesakoivu. This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye) and 825153 (EMBEDDIA).

References

1. Alokaili, A., Aletras, N., Stevenson, M.: Automatic generation of topic labels. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1965–1968 (2020)
2. Basave, A.E.C., He, Y., Xu, R.: Automatic labelling of topic models learned from twitter by summarisation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 618–624 (2014)
3. Bhatia, S., Lau, J.H., Baldwin, T.: Automatic labelling of topics with neural embeddings. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 953–963 (2016)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
5. Hulpus, I., Hayes, C., Karnstedt, M., Greene, D.: Unsupervised graph-based topic labelling using dbpedia. In: Proceedings of the sixth ACM international conference on Web search and data mining. pp. 465–474 (2013)
6. Kanerva, J., Ginter, F., Miekka, N., Leino, A., Salakoski, T.: Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics (2018)
7. Kim, H.H., Rhee, H.Y.: An ontology-based labeling of influential topics using topic network analysis. Journal of Information Processing Systems **15**(5), 1096–1107 (2019)
8. Kou, W., Li, F., Baldwin, T.: Automatic labelling of topic models using word vectors and letter trigram vectors. In: AIRS. pp. 253–264. Springer (2015)
9. Lau, J.H., Grieser, K., Newman, D., Baldwin, T.: Automatic labelling of topic models. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp. 1536–1545 (2011)
10. Lau, J.H., Newman, D., Baldwin, T.: Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 530–539 (2014)
11. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880 (2020)
12. Li, Y., Nair, P., Wen, Z., Chafi, I., Okhmatovskaia, A., Powell, G., Shen, Y., Buckeridge, D.: Global surveillance of covid-19 by mining news media using a multi-source dynamic embedded topic model. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. pp. 1–14 (2020)

13. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics **8**, 726–742 (2020)
14. Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L., Tolonen, M.: Topic modelling discourse dynamics in historical newspapers. arXiv preprint arXiv:2011.10428 (2020)
15. Maurya, K.K., Desarkar, M.S., Kano, Y., Deepshikha, K.: ZmBART: An unsupervised cross-lingual transfer framework for language generation. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 2804–2818. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.248, https://aclanthology.org/2021.findings-acl.248
16. Mei, Q., Shen, X., Zhai, C.: Automatic labeling of multinomial topic models. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 490–499 (2007)
17. Mueller, H., Rauh, C.: Reading between the lines: Prediction of political violence using newspaper text. American Political Science Review **112**(2), 358–375 (2018)
18. Popa, C., Rebedea, T.: BART-TL: Weakly-supervised topic label generation. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 1418–1425 (2021)
19. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
20. STT: Finnish news agency archive 1992-2018, source (http://urn.fi/urn:nbn:fi:lb-2019041501) (2019)
21. STT, Helsingin yliopisto, Alnajjar, K.: Finnish News Agency Archive 1992-2018, CoNLL-U, source (http://urn.fi/urn:nbn:fi:lb-2020031201) (2020), http://urn.fi/urn:nbn:fi:lb-2020031201
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
23. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating Text Generation with BERT. In: International Conference on Learning Representations (2019)