Interesting cross-border news discovery using cross-lingual article linking and document similarity

Boshko Koloski
Jožef Stefan Institute
Jožef Stefan IPS
Jamova 39, Ljubljana
boshko.koloski@ijs.si

Elaine Zosa
University of Helsinki
Pietari Kalmin katu 5, Helsinki
elaine.zosa@helsinki.fi

Timen Stepišnik-Perdih
Jožef Stefan Institute
Jamova 39, Ljubljana

Blaž Škrlj
Jožef Stefan Institute
Jamova 39, Ljubljana
blaz.skrlj@ijs.si

Tarmo Paju
Ekspress Meedia, Estonia
Narva mnt 13, Tallinn
tarmo.paju@delfi.ee

Senja Pollak
Jožef Stefan Institute
Jamova 39, Ljubljana
senja.pollak@ijs.si

Abstract

Contemporary news media face increasing amounts of available data that can be of use when prioritizing, selecting and discovering new news. In this work we propose a methodology for retrieving interesting articles in a cross-border news discovery setting. More specifically, we explore how a set of seed documents in Estonian can be projected in Latvian document space and serve as a basis for discovery of novel interesting pieces of Latvian news that would interest Estonian readers. The proposed methodology was evaluated by Estonian journalist who confirmed that in the best setting, from top 10 retrieved Latvian documents, half of them represent news that are potentially interesting to be taken by the Estonian media house and presented to Estonian readers.

1 Introduction

This paper presents our results of the participation in the hackaton, which was organised as part of the EACL 2021 Hackashop on news media content analysis and automated report generation. We are addressing the EMBEDDIA hackathon challenge on Identifying Interesting News from Neighbouring Countries (Pollak et al., 2021) in Estonian and Latvian context, which is a fully novel document retrieval task performed on recently released EMBEDDIA news datasets. Estonian journalists are very interested in identifying stories from Latvia, which will attract a large number of readers and are "special". While performing keyword-based search for Latvian news, where Estonians are mentioned is a simple task, this challenge on the contrary aims to identify a small set of documents from a larger number of topics, e.g. scandals, deaths and gossip that might be somehow connected to Estonia: not only by mentioning Estonians but by identifying news and stories that Estonians relate to (for example, when similar things have happened in Estonia or when similar news have been popular in Estonia).

In our approach, we first automatically create a collection of interesting articles using a string-based search and cross-lingual document linking, and then rank the query documents based on the proportion of interesting documents in their neighbourhood (where the neighbourhood is defined by a document similarity) by the newly introduced *Seed news of interest score (SNIR).*

The article first presents the datasets (Section 2), introduces the methodology (Section 3), and presents our experimental results (Section 4). The code and the data are made publicly available (see Section 5). Finally, Section 6 concludes the paper and presents the ideas for further work.

2 Datasets

In this study, we used the following resources.

- Archives of Estonian news articles from Ekspress Meedia. Ekspress Meedia belongs to Ekspress Meedia Group, one of the largest media groups in the Baltics. From the entire collection of Ekspress Meedia articles (Pollak et al., 2021), we selected the articles from the years 2018 and 2019 (i.e. 64,651 articles in total).

- Dataset of archives of Latvian news articles (Pollak et al., 2021) come from Latvian Delfi that also belongs to Ekspress Meedia Group. We considered only the articles from the years 2018 and 2019 (i.e. 60,802 articles in total).

Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 116–120
April 19, 2021 © Association for Computational Linguistics

- Manually identified interesting news for Estonian readers in Latvian (and their Estonian counterparts). These were manually identified as examples of interesting news by Estonian journalist from Ekspress Meedia. Note that the Estonian articles are not their direct translations, as the articles can be slightly adapted to Estonian audience.

3 Methodology

Our methodology consists of two steps. First, we automatically construct the datasets of interesting Latvian articles and next propose a method to retrieve interesting articles by ranking a given query document based on the the proportion of interesting articles in its neighbourhood.

3.1 Automated selection of Latvian example articles

The aim of this step is to automatically construct *Latvian seed news of interest*, which are considered as good examples of interesting Latvian articles. As there are only 21 manually identified examples, which we keep for the evaluation purposes and parameter setting, this step was automatised.

In our approach, we first extract Estonian articles, that specifically mention the source of Latvian Delfi (*Läti Delfi*), which leads to 100 identified Estonian articles which are considered as automatically constructed Estonian example data. Then we follow the methodology of Zosa et al. (2020). More specifically, we use Sentence-BERT (Reimers and Gurevych, 2019) to obtain cross-lingual encodings of the articles from both languages. For each Estonian article, we extract best *k* Latvian candidates (where *k* is a parameter) by taking the cosine similarities between the query Estonian article and all the candidate Latvian articles and finally rank the Latvian articles based on this similarity measure.

Note that also recent work (Litschko et al., 2021) has shown that specialized sentence encoders trained for semantic similarity across languages obtain better results in document retrieval than static cross-lingual word embeddings or averaged contextualized embeddings.

3.2 Retrieval of interesting news articles

In this step, we assign the ”interestingness score” to each query article. First, we identify the local neighbourhood of a query article by document similarity. We use the same sentence-embeddings method as in the previous step, with the difference that here the articles similarity is computed in monolingual setting. The number of articles surrounding the query is a parameter *m*.

We introduce a custom metric called SNIR (*Seed news of interest ratio*), where we compute the ratio of automatically identified interesting news compared to all the articles in the neighbourhood. The hypothesis is that the articles of interest will have more articles from the automatically identified interesting news articles in their surrounding than the articles, which are not relevant for the Estonian readers.

The result of our method is a ranked list of articles for a given time period (e.g. a day, week, month) where a journalist can then decide to manually check top *x* articles. In addition, in future also a SNIR threshold could be set which would allow interested journalists to be informed about potentially interesting articles in real-time.

The SNIR score is defined as follows. Let $\text{NeighborhoodDocuments}_m$ represent the set of m nearest documents in the final embedding space. Let Interesting_m represent the set of m *interesting* seed documents obtained via the cross-lingual mapping discussed in the previous sections. We can define the SNIR at m as:

$$\text{SNIR}(m) = \frac{|\text{Interesting}_m|}{|\text{NeighborhoodDocuments}_m|}.$$

We report SNIR values for different neighborhood (m) sizes. The goal of SNIR is to score interesting query articles higher than query articles which are not of special interest for Estonian readers.

4 Experiments and results

4.1 Automated analysis

For our experiments, we used the following settings. We test the parameter setting for *k* in cross-lingual article linking to 20 and 100, and the setting of parameter *m* to 10, 20 and 100 for determining the neighbourhood in computing the SNIR score.

First, we evaluated the cross-lingual article linking on the 21 manually linked article pairs. For these article pairs, We obtained an MRR (Mean Reciprocal Rank) score of 64.93%, which shows that for an article in a source language, the correct article is usually proposed as the first or second candidate.

Next, we performed qualitative analysis by visualising the document space. In Figure 1 (using parameter k=20), we can see that automatically defined Latvian seed news of interest (red) are not evenly distributed and support the hypothesis, that random article's neighbourhood will differ in this respect. The figure also presents the manually identified interesting news (orange), where at least some of the documents seem to be positioned together.

Next, we compare the SNIR scores of 21 manually identified interesting articles compared to random Latvian articles. The results of SNIR score for parameter *k*=100 at different *m* can be found in Figure 2. This also suggests that there is some evidence that a threshold could in future be determined, but more extensive experiments should be performed in future work.

4.2 Manual analysis

For final evaluation, we selected the last month of the Latvian collection (1408 articles in total), and ranked the articles according to the SNIR score. These were provided to the Estonian journalist of Ekspress Meedia who evaluated top 10 results for each of the settings.

We prepared four different pairs of $k \in 10, 100$ retrieved documents and $m \in 10, 20, 100$ documents in the neighbourhood which were evaluated by the media expert. The media house expert analyzed the retrieved documents and labeled them with three different labels based on the acceptance:

- No - the article was of *not relevant* significance to the media house.
- Maybe - the article contained news about events that *are potentially relevant* to the Estonian readers.
- Yes - the article contained news about events that *are relevant* to the Estonian readers or contained extraordinarily news.

The evaluation of the top 10 articles retrieved for each k, m pair is listed in Table 1.

From the evaluation we can see that when we have a relatively small number of retrieved documents and a smaller neighbourhood we can benefit from the SNIR metric. As the best performing parameter pairs were the $k = 20$ and $m = 10$ retrieving 50% articles as relevant or of close relevance to the Estonian news house. When larger neighbourhood is introduced the space becomes sparser and the method tends to retrieve more false positives.

k	m	Yes	Maybe	Not
20	10	2	3	5
20	100	1	3	6
100	20	1	3	6
100	100	0	3	7

Table 1: Evaluation of the top-10 retrieved articles by the SNIR ranking for various k interesting Latvian seeds documents and m neighbourhood sizes.

The journalist also explained why a selected news example from positive category is very relevant. The news talks about a scooter accident in court proceedings, which is *extraordinary*, as well as *relevant to Estonians* as the debate around scooters at the streets is also very active in Estonia. Some examples from negative category contain articles about foreign news (terror attack, for example) and these are not the type of news that the Estonian journalists would pick from Latvian media.

5 Availability

The code and data of the experiments is made available on the GitHub: `https://github.com/bkolosk1/Interesting-cross-border-news-discovery`

6 Conclusion and future work

In this work we tackled the problem of retrieving interesting news from one country for the context of another neighbouring country. We focused on finding interesting news in Latvian news space that would be engaging for the Estonian public. We used Latvian and Estonian EMBEDDIA datasets to construct the document space. First we used a string matching approach to identify a subset of news in Estonian media that originated from Latvian news. Next, we utilized the methods for *ad hoc* Cross Lingual document retrieval to find corresponding articles in the Latvian news space. After automatically retrieving this set of Latvian news articles of interest, we used this information in a novel metric defined as SNIR, that analyses a news article's neighbourhood in order to measure it's relevance (interestingness). The assumption of the metric is that if the surrounding documents of a query point are relevant, this new point might be of relevance. The SNIR scores of randomly selected 20 documents and 20 documents identified as examples of interesting news by an Estonian journalist showed that their value differ, which is

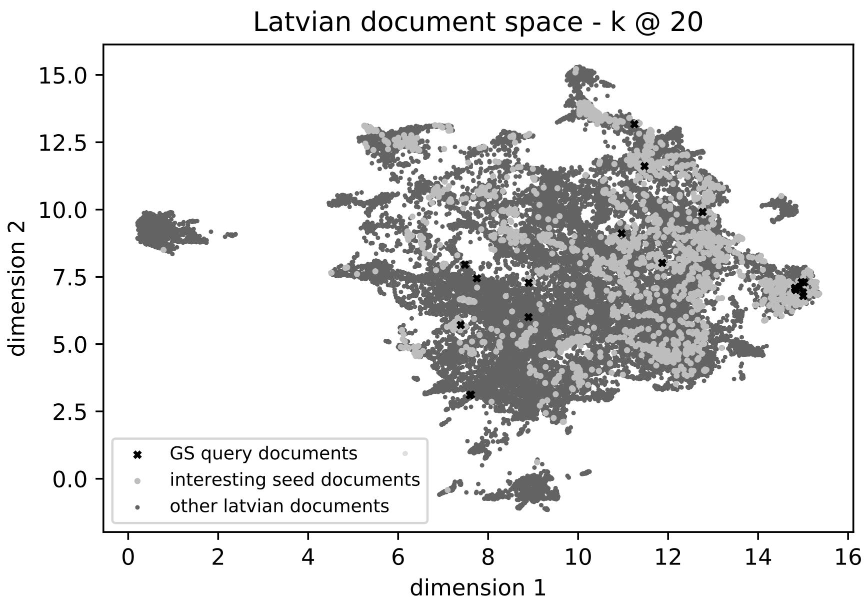

Figure 1: MAP 2D projection of the Latvian data, where black crosses mark query docs (GS) represent gold standard, i.e. manually identified Latvian news of interest to Estonian readers, gray dots represent automatically identified Latvian seeds (identified by string-based search in Estonian and cross-lingual linking to Latvian) and dark-gray dots represent all other Latvian documents.

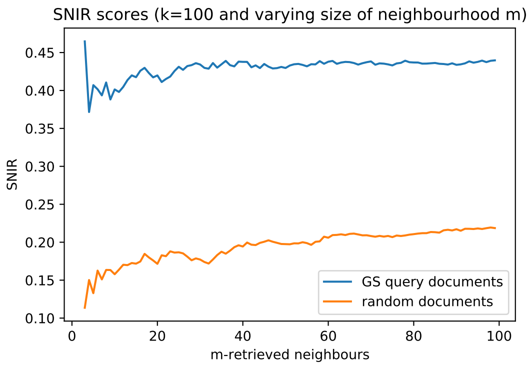

Figure 2: Evaluation of the SNIR metric for the 21 gold standard queries (manually identified news of interest) and 21 random query points. The results indicate that a random documents' neighbourhood is structured differently compared to the one of relevant interesting documents.

promising. Finally, we prepared a test set of news from one month and sent them to manual evaluation by a journalist. Results of top 10 candidates of each setting suggest that the proposed metric works well if the parameters of interesting articles and neighborhood were adjusted right, with the best performing parameter tuple yielding 50% hit-ratio.

For the further work we propose exploring the keywords appearing in the clusters of interesting news and exploiting their named entity tags in order to achieve even better performance. We also want to include background knowledge from knowledge graphs to improve the document similarity evaluation. Special attention will also be paid to setting a threshold for SNIR which would allow for real-time investigation of best candidates in a real journalistic practice.

7 Acknowledgements

The work was supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The work was also supported by the Slovenian Research Agency (ARRS) through core research programme *Knowledge Technologies* (P2-0103) and the work by B.Š. through the ARRS young researcher grant.

References

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2021. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In *Proceedings of ECIR*.

Senja Pollak, Marko Robnik Šikonja, Matthew Purver, Michele Boggia, Ravi Shekhar, Marko Pranjić, Salla

Salmela, Ivar Krustok, Tarmo Paju, Carl-Gustav Linden, Leo Leppänen, Elaine Zosa, Matej Ulčar, Linda Freienthal, Silver Traat, Luis Adrián Cabrera-Diego, Matej Martinc, Nada Lavrač, Blaž Škrlj, Martin Žnidaršič, Andraž Pelicon, Boshko Koloski, Vid Podpečan, Janez Kranjc, Shane Sheehan, Emanuela Boros, Jose Moreno, Antoine Doucet, and Hannu Toivonen. 2021. EMBEDDIA tools, datasets and challenges: Resources and hackathon contributions. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Elaine Zosa, Mark Granroth-Wilding, and Lidia Pivovarova. 2020. A comparison of unsupervised methods for ad hoc cross-lingual document retrieval. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 32–37, Marseille, France. European Language Resources Association.