



EMBEDDIA

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

Research and Innovation Action

Call: H2020-ICT-2018-1

Call topic: ICT-29-2018 A multilingual Next generation Internet

Project start: 1 January 2019

Project duration: 39 months

D4.8: Final evaluation report on cross-lingual content analysis technology (T4.4)

Executive summary

WP4 focused on analysis of news content in form of news articles across languages. This deliverable presents the final results of T4.4 “Resource gathering, benchmarking and evaluation”. It introduces the datasets that were released by the EMBEDDIA project, including unannotated datasets and annotated datasets, and presents final evaluation of selected tools of WP4 on the media partners’ datasets as well as on the public datasets (e.g. in the scope of participation to the shared tasks).

Partner in charge: TEXTA

Project co-funded by the European Commission within Horizon 2020 Dissemination Level

PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	–
RE	Restricted to a group specified by the Consortium (including the Commission Services)	–
CO	Confidential, only for members of the Consortium (including the Commission Services)	–



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

Deliverable Information

Document administrative information	
Project acronym:	EMBEDDIA
Project number:	825153
Deliverable number:	D4.8
Deliverable full title:	Final evaluation report on cross-lingual content analysis technology
Deliverable short title:	Final evaluation of content analysis technology
Document identifier:	EMBEDDIA-D48-FinalEvaluationOfContentAnalysisTechnology-T44-submitted
Lead partner short name:	TEXTA
Report version:	submitted
Report submission date:	28/02/2022
Dissemination level:	PU
Nature:	R = Report
Lead author(s):	Linda Freienthal (TEXTA), Senja Pollak (JSI)
Co-author(s):	Andraž Pelicon (JSI), Boshko Koloski (JSI), Blaž Škrlić (JSI), Matej Martinc (JSI), Adrian Cabrera (ULR), Emanuela Boros (ULR), Elaine Zosa (UH), Lidia Pivovarovova (UH), Ivar Krustok (ExM)
Status:	_ draft, _ final, <u>x</u> submitted

The EMBEDDIA Consortium partner responsible for this deliverable has addressed all comments received. Changes to this document are detailed in the change log table below.

Change log

Date	Version number	Author/Editor	Summary of changes made
24/01/2022	v0.1	Linda Freienthal (TEXTA)	Produced first draft
27/01/2022	v0.2	Elaine Zosa (UH)	Topic labelling using ontology mapping
01/02/2022	v1.0	E. Boros and A. Cabrera (ULR)	TREC description
10/02/2022	v1.1	Matej Martinc (JSI), Lidia Pivovarovova (UH), Linda Freienthal (TEXTA)	Russian diachronic semantic change and TEXTA evaluation
11/02/2022	v1.3	Blaž Škrlić (JSI)	AutoBot evaluation
12/02/2022	v1.4	E. Zosa (UH), E. Boros (ULR)	SemEval News similarity description
13/02/2022	v1.5	Andraž Pelicon (JSI)	Sentiment evaluation
16/02/2022	v1.6	Boshko Koloski (JSI)	SemEval and Interesting news discovery
17/02/2022	v1.7	Senja Pollak (JSI)	Final draft
23/02/2022	v2.0	Hannu Toivonen (UH)	Internal review
23/02/2022	v2.1	Matthew Purver (QMUL)	Internal review
24/02/2022	v2.2	all	Changes after internal review
26/02/2022	v2.3	Nada Lavrač (JSI)	Quality control
27/02/2022	final	Linda Freienthal (TEXTA)	Final report
28/02/2022	submitted	Tina Anžič (JSI)	Report submitted

Table of Contents

1. Introduction.....	5
2. Publicly released media partners' datasets	5
2.1 Unannotated datasets	5
2.2 Annotated datasets	7
3. Evaluation of EMBEDDIA tools on media partners' datasets.....	8
3.1 Evaluation of keyword extraction	8
3.1.1 Evaluation by ExM	8
3.1.2 TEXTA evaluation on the data of the National Library of Estonia	9
3.2 Evaluation of sentiment analysis	9
3.2.1 Data	9
3.2.2 Results	10
3.3 Evaluation of cross-lingual interesting news retrieval.....	11
3.3.1 Data	11
3.3.2 Method	11
3.3.3 Classification settings	12
3.3.4 Threshold learning	12
3.3.5 Evaluation	13
4. Final evaluation of EMBEDDIA tools on public datasets.....	14
4.1 Evaluation of autoBOT for the task of fake news classification against a generation of computer science students.....	14
4.2 RuShiftEval: a shared task on semantic change detection for Russian.....	15
4.2.1 Dataset creation.....	15
4.2.2 Task formulation for the shared task	17
4.2.3 Results	17
4.3 TREC 2021: News Track background linking task	18
4.3.1 Data	19
4.3.2 Data indexing	19
4.3.3 Background linking approaches	19
4.3.4 TREC 2021 background linking results	22
4.4 SemEval 2022 Task 8: Multilingual News Similarity.....	23
4.4.1 Data	23
4.4.2 Approaches for assessing the similarity between news	23
4.4.3 Semantic textual similarity models.....	23
4.4.4 Image similarity and regression models	25
4.4.5 Knowledge graph similarity and regression models	26
4.4.6 Text and image regression models	27
4.5 Multilingual Topic Labelling of News Topics using Ontological Mapping	28
4.5.1 Models.....	29
4.5.2 Datasets	29
4.5.3 Results	29
4.5.4 Conclusions	31
5. Conclusions	31
6. Associated Outputs	32

References	34
Appendix A: EMBEDDIA Tools, Datasets and Challenges: Resources and Hackathon Contributions..	41
Appendix B: RuShiftEval: a shared task on semantic shift detection for Russian	52
Appendix C: Three-part diachronic semantic change dataset for Russian.....	66
Appendix D: Multilingual Topic Labelling of News Topics using Ontological Mapping	73
Appendix E: Elastic Embedded Background Linking for News Articles with Keywords, Entities and Events.....	81

List of abbreviations

D	Deliverable
EMA	EMBEDDIA Media Assistant
ExM	Ekspress Meedia Group
IPTC	International Press Telecommunications Council
KG	Knowledge graph
NE	Named Entity
STT	Finnish News Agency
T	Task
TRI	Trikoder

1 Introduction

WP4 dealt with analysis of news content across languages to overcome the language barriers and overflow of information. In contrast to WP3, where the focus is on short texts (news comments), WP4 focused primarily on news content in the form of articles.

Results of Tasks T4.1–T4.3, described in previous deliverables of WP4, contained information on the tools for linking of relevant texts, summarisation and visualisations of content, and analysis of the viewpoints and sentiment of articles from different sources.

In this report we present the final results of Task T4.4., which focused on gathering and preprocessing training and testing data provided by our media partners and using them to evaluate our EMBEDDIA tools created within WP4. We released a large number of datasets and tools and proposed a number of challenges in the scope of the Hackashop on News Media Content Analysis and Automated Report Generation that we organised in conjunction with EACL and which took place during a three-week period in February 2021 (Pollak, Robnik-Šikonja, et al., 2021). Next, we enriched the media partners' datasets with annotations from Tasks T2.1 and T2.2. These datasets were also used for evaluating the tools developed within the work package (such as keyword extraction, sentiment analysis and cross-lingual interesting news retrieval). In addition, we continued the development of methods, which were evaluated on a number of public datasets, especially in the scope of shared tasks such as RuShiftEval (Schlechtweg et al., 2020), TREC 2021¹ and SemEval 2022.

This deliverable gives an overview of the datasets produced (Section 2), results of the evaluations done on media partners' datasets (Section 3), and results of the evaluations performed on public datasets (Section 4). It also presents five papers, which are added as appendices, covering (1) EMBEDDIA Tools and datasets (Appendix A), (2) a first shared task on diachronic word meaning change detection in Russian (Appendix B) and a dataset for it (Appendix C), as well as (3) a new method for multilingual topic labelling (Appendix D), and several methods for news background linking (Appendix E).

2 Publicly released media partners' datasets

In this section, we present the datasets that we released publicly during the EMBEDDIA project. Many of the datasets and tools were for the first time shared with the public in April 2021, in the scope of the hackashop (hackathon+workshop) that we organised at the EACL 2021 conference. We prepared it for the the hackathon part, where we brought together about 25 active hackathon participants and about 20 researchers from EMBEDDIA, who could choose to work with EMBEDDIA data and tools or the ones of their own interest. This was a joint effort across different EMBEDDIA WPs, but as many of the datasets and tools concern WP4, we report it in this deliverable as Appendix A (Pollak, Robnik-Šikonja, et al., 2021) where the details of the data can be found. Below, we first summarise the unannotated datasets (Section 2.1). After the hackashop we also produced annotated versions which are described next (Section 2.2).

2.1 Unannotated datasets

In Deliverable D4.1, we described some of the datasets by media partners, but at that time they were made available only for internal use in the project. In this deliverable, we present the public releases where datasets are available to the wider public through CLARIN and other similar repositories. Unannotated datasets are described also in paper of Appendix A (Pollak, Robnik-Šikonja, et al., 2021), and briefly summarised below.

The unannotated media partners' datasets made public are as follows:

¹<http://trec-news.org/>

- **Ekspress Meedia News Archive (in Estonian and Russian) 1.0** (Purver, Pollak, et al., 2021). Ekspress Meedia belongs to the Ekspress Meedia Group, one of the largest media groups in the Baltics. This dataset has over 1.4M articles from Ekspress Meedia news site from 2009-2019, mostly in the Estonian (1,115,120 articles) with some in the Russian language (325,952 articles). Keywords (tags) are included for articles after 2015. The dataset is publicly available in the CLARIN repository.²
- **Latvian Delfi article archive (in Latvian and Russian) 1.0** (Pollak, Purver, et al., 2021). Latvian Delfi belongs to Ekspress Meedia Group. This dataset contains over 180,000 articles from Delfi news site, half of them being in Latvian and half of them in Russian. Keywords (tags) are included. The data is publicly available in the CLARIN repository.³
- **24sata news article archive 1.0** (Purver, Shekhar, et al., 2021). 24sata is the biggest Croatian news publisher, owned by the Styria Media Group. The dataset contains over 650,000 articles in Croatian between 2007–2019, as well as assigned tags. The data is publicly available in the CLARIN repository.⁴
- **Finnish news agency archives.** Three versions of the data of Finnish news agency STT were released. First, *Finnish News Agency Archive 1992-2018* (STT, 2019) is the Finnish News Agency Archive corpus comprising newswire articles in Finnish sent to media outlets by the Finnish News Agency (STT) between 1992-2018 and is made available through MetaShare⁵. The dataset is also available in CONLL-U format as resource *Finnish News Agency Archive 1992-2018, CoNLL-U*, (STT et al., 2020), also on MetaShare⁶. Finally, a more recent version of the archive was released, *Finnish News Agency Archive 2019-2021* (STT, 2022), available through MetaShare.⁷

We also released a selection of task-specific datasets:

- **Keyword extraction datasets for Croatian, Estonian, Latvian and Russian 1.0** (Koloski et al., 2021b) were released in collaboration with WP2. They were created for keyword extraction tasks and presented in the hackashop. The language distributions are follows:
 - Croatian: 32,223 train, 3,582 test;
 - Estonian: 10,750 train, 7,747 test;
 - Russian: 13,831 train, 11,475 test;
 - Latvian: 13,133 train, 11,641 test.

It contains the tags added by the editors of participating media houses. The datasets are available in CLARIN⁸. The datasets were used for development and evaluation of keyword extraction systems (Koloski et al., 2021a; Koloski, Pollak, et al., 2022).

- **Sentiment Annotated Dataset of Croatian News** (Pelicon et al., 2020b). This is a subset from the Croatian 24sata news archive (see above) annotated with manually annotated sentiment scores. The annotation guidelines were presented in Deliverable D4.1, and experiments where these datasets were used in experiments presented in Deliverable D4.7 and in (Pelicon et al., 2020a).
- **Estonian-Latvian Interesting News Pairs.**⁹ These are manually identified interesting news for Estonian readers from Latvian news media (and their Estonian counterparts). These were manually identified as examples of interesting news by Estonian editor from Ekspress Meedia. Note that

²<http://hdl.handle.net/11356/1408>

³<http://hdl.handle.net/11356/1409>

⁴<http://hdl.handle.net/11356/1410>

⁵<http://urn.fi/urn:nbn:fi:lb-2019041501>

⁶<http://urn.fi/urn:nbn:fi:lb-2020031201>

⁷<https://metashare.csc.fi/repository/browse/finnish-news-agency-archive-2019-2021-source/ee6145c2882211eca1f5fa163ec5ae3e1d0fa3d38e314897bb2e5cdcf0fa021b/>

⁸<http://hdl.handle.net/11356/1403>

⁹<https://github.com/EMBEDDIA/interesting-cross-border-news-discovery>

the Estonian articles are not their direct translations, as the articles can be slightly adapted to Estonian audience. The dataset was created for the challenge and approach on finding interesting news from neighbouring countries, described by (Koloski, Zosa, et al., 2021) and in Section 3.3.

A more detailed description of unannotated datasets and selected tools are described in (Pollak, Robnik-Šikonja, et al., 2021) attached here as Appendix A.

2.2 Annotated datasets

We annotated a sample corpus of Estonian, Croatian and Latvian news articles with EMBEDDIA tools and published this corpus at CLARIN¹⁰ (Freienthal et al., 2022). The purpose of this dataset is to make our tools' results available for analysis and usage.

This dataset contains the following collections of articles from EMBEDDIA Media partners:

- 12,390 Estonian articles from 2019 (including original tags given by Ekspress Meedia), which is a subset of the unannotated dataset by ExM (Purver, Pollak, et al., 2021),
- 5,000 Croatian articles from autumn of 2010 (including original tags given by 24sata), which is a subset of the unannotated dataset by 24sata (Purver, Shekhar, et al., 2021),
- 15,264 Latvian articles from 2019 (including original tags given by DELFI), which is a subset of the unannotated dataset by Delfi from Ekspress Meedia Group (Pollak, Purver, et al., 2021).

The articles in the dataset have been annotated with the following EMBEDDIA tools, after preprocessing them with texta-mlp Python package¹¹ via the EMBEDDIA Media Assistant's Texta Toolkit:¹²

- **Named Entity Recognition Tool modules Latin1 and Latin2** (Cabrera-Diego, Moreno, & Doucet, 2021): Names of people, organizations, and locations are called named entities (NEs). These are often the most important pieces of information people search for in articles and can be automatically extracted. The tools Latin1 and Latin2 that were used to extract NEs with their label (e.g. PER as persona), value (e.g. "Johnny Depp") and span (place in text) in this dataset were made also available in HuggingFace: <https://huggingface.co/creat89>.
- **RaKUn keyword extractor.** RaKUn (Škrlj et al., 2019) is an unsupervised system for keyword extraction, so it can be used for any language. It produces annotations in the form of keyword-score tuples, where keywords are a single- or multi-term phrases present in a given document. Keywords extracted with RaKUn are added to the article with an extra field in the JSON-lines document.
- **TNT-KID keyword extractor.** TNT-KID (Martinc et al., 2021) is a supervised system for automatic keyword extraction. It was trained on a corpus of articles with human-assigned keywords. For Croatian, the annotators were 24sata editors, for Estonian the Ekspress Meedia staff and for Latvian the Latvian Delfi staff. For Croatian only TNT-KID was applied, while for Estonian and Latvian, the TNT-KID with TF-IDF, and extension by (Koloski et al., 2021a) was used. Keywords extracted by TNT-KID were added to the article with an extra field in the JSON-lines document.
- **Sentiment analysis.** Our news sentiment analyser (Pelicon et al., 2020a) labels a news article as being of positive, negative, or neutral sentiment, using a fine-tuned multilingual BERT model, which was trained on Slovene sentiment annotated news articles.

All the data is encoded in "JSON-lines" format.

¹⁰<http://hdl.handle.net/11356/1485>

¹¹<https://pypi.org/project/texta-mlp/>

¹²<https://docs.texta.ee/>

3 Evaluation of EMBEDDIA tools on media partners' datasets

We evaluated our two keyword extraction methods, a sentiment analysis method and a cross-border news extractor method on our media partners' datasets to estimate how our tools work on real-life data. We present the results in this chapter.

3.1 Evaluation of keyword extraction

In Deliverables 2.3 and 2.6 we introduced methods for extracting terms and keywords from the input text in a monolingual and multilingual problem setting. In this section we evaluate the results of the graph-based key-word extraction method RaKun (Škrlić et al., 2019) and a Transformer-Based Neural Tagger for Keyword IDentification called TNT-KID (Martinc et al., 2021) on Estonian articles.

3.1.1 Evaluation by ExM

Ekspress Meedia publishes news on several subsites/subpages. For ensuring that the data set covers the entire variety of news genres, we divided the news into 5 topics and got 20 articles from each of the topic (total 100 articles):

- Entertainment. Articles from <https://kroonika.delfi.ee/>.
- Business. Articles from <https://arileht.delfi.ee> and <https://epl.delfi.ee>.
- Express. Articles from <https://ekspress.delfi.ee>.
- Varia. Articles from several different subpages such as <https://kinoveeb.delfi.ee> and <https://moodnekodu.delfi.ee>.
- Magazine. Articles from several different subpages such as <https://omamaitse.delfi.ee> and <https://tervispluss.delfi.ee>.

We divided the dataset into topics, because we wanted to make sure that the evaluators from Ekspress Meedia (news journalists and editors) evaluate keywords on articles they usually work with. That means that the business news editor didn't have to evaluate entertainment news etc.

Each article was evaluated by two annotators. The annotator looked at the article, the keywords given to the article before by human annotators, and the results of keywords assigned by TNT-KID and RaKun. The annotators then marked down those of the above keywords he/she would use, keywords he/she would add, and keywords that were completely off and shouldn't be used. They also gave their opinion of the outputs of the tools using scale 1–5, where:

- 1 means that the keywords are not relevant to the article at all and don't give proper overview of the content
- 2 means that a few of the keywords are relevant to the article, but the keywords in total give a wrong idea of the content.
- 3 means that there are keywords relevant to the article, but also many irrelevant keywords that do not provide a completely clear idea of the content.
- 4 means that in overall the keywords are relevant to the article, only some of them are misleading.
- 5 means that the keywords are relevant to the content and give a proper idea/overview of it.

The results showed 1.14 points for RaKun and 2.43 for TNT-KID. Since the RaKun hyperparameters were not configured for Estonian text and it wasn't run on lemmatized text, the evaluators evaluated it

quite poorly with the average being 1.14 points. Although TNT-KID also wasn't run on lemmatized text, it got better results with the average being 2.43. A majority of the tags marked as missing from the output of the methods were names of people, which means that in future potentially TNT-KID and named entity results should be combined.

From the perspective of ExM, the results of TNT-KID are satisfactory. The system performs better than the current solution used by ExM, and it is an appropriate solution for the implementation in the life product.

We also note that in TNT-KID training, the datasets were much older than the data used in evaluation. This can also be one of the potential differences between quantitative, automated evaluation reported in deliverables of WP2 and the manual evaluation here. For optimal results, We recommend that one should regularly (e.g. twice a year) update the system.

3.1.2 TEXTA evaluation on the data of the National Library of Estonia

TEXTA's own testing with RaKUn revealed that with appropriate preprocessing (lemmatization) and hyperparameters the results were good enough with Estonian texts for production value. Therefore TEXTA added RaKUn to Texta Toolkit.¹³

In fact, TEXTA tested RaKUn in the Texta Toolkit as one of the methods for solving automatic subject indexing (this means keyword tagging) in National Library of Estonia. Documentation about the tender in Estonian can be found here: <https://riigihanked.riik.ee/rhr-web/#/procurement/3224632/documents?group=B>.

Both regular library users and the library's cataloguers tested 7 different methods. The evaluation showed that RaKUn was the best method out of all the methods tested out in this tender. See Figure 1 for an example of the tender's prototype¹⁴. The results of the evaluation will be available in a detail analysis which will be the outcome of the tender and will be published in the tender's page referred to above. Currently the analysis is not yet public.

Given the speed of the method, as well as relatively good precision with right parameter setting, TEXTA will continue to use RaKUn for other similar problems, as integrated into the Texta Toolkit.

3.2 Evaluation of sentiment analysis

In Deliverable D4.7 we introduced the task of monolingual and cross-lingual identification of viewpoints and sentiment in news reporting. In this section, we present an evaluation of sentiment analysis (Pelicon et al., 2020a) on Estonian articles (Purver, Pollak, et al., 2021).

3.2.1 Data

We randomly selected 100 articles in Estonian from our media partner Ekspress Media's dataset (Purver, Pollak, et al., 2021). The articles were then annotated with labels "neutral", "positive" and "negative" by two annotators. Their guideline was to answer the question "Did this news evoke positive/neutral/negative negative feelings?". The annotators did not agree in 45 cases. These cases were solved by third annotator who decided between the two chosen options.¹⁵

¹³It was added in version 2.41. Documentation about Texta Toolkit version 2.10 can be found here: docs.texta.ee

¹⁴The example article was taken from here: <https://www.hm.ee/et/uudised/wiedemanni-keeleauhinna-palvis-eesti-keeletehnoloogia-rajaja-mare-koit>

¹⁵For the two annotators, the inter-annotator agreement (IAA) was computed using Krippendorff Alpha. For Estonian, the Kalpa metric is 0.335. For reference, the Kalpa metric for the Croatian test set is 0.441 and for the Slovenian dataset is 0.454.

Kratt: Artiklite automaatne märksõnastaja prototüüp

Artikli valimine

Tekst

Fail

URL

Sisesta tekst...



Wiedemanni keeleauhinna pälvis Eesti keeletehnoloogia rajaja Mare Koit

Wiedemanni keeleauhinna laureaati Mare Koit. Foto: Tartu Ülikool. Valitsus otsustas määrata 2022. aasta riigi F. J. Wiedemanni keeleauhinna Mare Koidule silmapaistva tegevuse eest Eesti keeletehnoloogia ja arvutilingvistika rajaja ning arendajana.

Haridus- ja teadusminister Liina Kersna õnnitles ja tänas laureaati. „Mare Koit on esimene keeletehnoloogia professor Eestis ning ta on jätkuvalt valdkonna üks mõjukamaid kujundajaid,” ütles Kersna. „Tänu Mare Koidu selgele visioonile ning katkematule tööle nii teadlase kui ka õppejõuna on Eesti keeletehnoloogia ka rahvusvahelises võrdluses väga kõrge tasemel.”

Mare Koit on teadlasena pühendunud tehisintellekti arendamisele, tagades kõik vajaliku selleks, et me

Staatust: OK

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti
tuleviku heaks

Sätted

Märksõnasta

Tuvastatud keeled: eesti

Peida kõik

Kuva seaded

Seotud märksõnad

Saada tagasiside

MARC

Teemamärksõnad

Leksikaalne 2/16

SUCCESS

RaKUn 5/25

SUCCESS

Minimaalne tõenäosus

Mare koit

Wiedemanni keeleauhind ✓

keeletehnoloogia ✓

eesti keel ✓

rajaja

Hybrid Tagger 3/23

SUCCESS

Figure 1: View of the automatic keyword tagging prototype for National Library of Estonian. Here we used news about Mare Koit, who got Wiedemann’s language prize (*Wiedemanni keeleauhind* in Estonian) for her outstanding work in and being a founder (*rajaja*) of Estonian (*eesti keel*) language technology (*keeletehnoloogia*) and Estonian computational linguistics. RaKun predicted keywords are opened in this figure. Checkmarks indicate that these keywords exist in the Estonian Subject Thesaurus (see <https://ems.elnet.ee/index.php>).

3.2.2 Results

Table 1: Results of EMBEDDIA zero-shot sentiment classifier by Pelicon et al. (2020a), trained on Slovenian articles evaluated on Estonian news articles, compared to the Majority class baseline.

	Accuracy	Recall	Precision	macro F1
Majority Baseline Classifier	0.40	0.33	0.13	0.19
Multilingual Sentiment Classifier	0.57	0.54	0.70	0.55

We compare the results of the multilingual sentiment model described in (Pelicon et al., 2020a), and of a simple majority-class classifier, to the golden dataset described above. The results are presented in Table 1. The accuracy and macro F1-score of the multilingual sentiment model are 0.57 and 0.55, respectively. Comparing it with the simple majority-class classifier, it performs substantially better. Additionally, the results are similar to the results on the Croatian data where the model was also tested in zero-shot setting, where macro F1 score was 54.77 (Pelicon et al., 2020a).

3.3 Evaluation of cross-lingual interesting news retrieval

In Deliverable D4.5, we introduced the task of retrieving and extracting interesting cross-border news, relevant for translations for media houses. In this section, we introduce a novel method, which uses auto-encoder neural architecture in order to extract the relevant documents.

3.3.1 Data

The data used in this work consists of Estonian and Latvian articles (published in the period between 01.01.2018 until 01.12.2019) by media houses belonging to the Ekspress Meedia Group. More specifically, from the EMBEDDIA news archives data set (Pollak, Robnik-Šikonja, et al., 2021), we used the following subcorpora:

- The collection of **Estonian** news articles from the archives of Ekspress Meedia, resulting in 17148 articles¹⁶.
- The collection of **Latvian** news articles published by the DELFI portal, a Latvian subsidiary of the Ekspress Meedia Group. Similarly to (Koloski, Zosa, et al., 2021), we use the data before December 1, 2019, for training (29178 articles), and the data after for testing (1339 articles).
- The set of **21** pairs of aligned **Estonian** and **Latvian** news, consisting of selected articles published (between January 1, 2019 and December 31, 2019) in the Latvian journal and their news counterparts adapted to the Estonian readers, manually retrieved by an Estonian journalist.

3.3.2 Method

Automated acquisition of Estonian ground truth. Similarly to (Koloski, Zosa, et al., 2021), our method consists of two steps. In the first step, we follow the approach from (Koloski, Zosa, et al., 2021) using exact string matching to extract Estonian articles that mention Latvian Delfi (Läti Delfi, Lati Delfi, Delfi.lv) in the article body text as a source of news. The hypothesis is that these articles were identified as of significance for translation/adaptation from their Latvian original counterparts. In this manner, we acquired 100 Estonian articles, we denote them as $Estonian_{ground}$.

Cross-lingual mapping. We hypothesize that potentially interesting Latvian news are the ones that appear closest to each Estonian article of the $Estonian_{ground}$ in the joint multilingual space. To do so, as in (Koloski, Zosa, et al., 2021), we follow the methodology in (Zosa et al., 2020) for extracting articles in a multilingual setting:

1. We use sentence-transformers (Reimers & Gurevych, 2019b) XLM-BERT-PASSPHRASE embeddings to embed the articles from $Estonian_{ground}$ and the $Latvian_{train}$ articles in a common, multilingual space.
2. For each article $E_i \in Estonian_{ground}$ collection, we select $k \in \{1, 100\}$ closest Latvian articles, obtaining a collection of Latvian articles $LE_{i,k}$ for each article E_i .
3. Finally, we join all of the sets $LE_{i,k}$ from the previous step, obtaining the final **Latvian_{extracted@k}** Latvian extracted set of articles. Formally: $Latvian_{extracted@k} = \bigcup_{i=1}^{100} LE_{i,k}$ for a given neighborhood parameter k .

To evaluate the mapping, the Mean Reciprocal Rank (MRR) between the mappings of Estonian to Latvian articles, and vice-versa, were computed on the 21 pairs, where we obtained an average MRR of 66.67%. Even if the linking is not always correct, we assume that even when we do not retrieve the exact match, the articles in the identified neighbourhood k still represent a neighbourhood of potentially

¹⁶The original dataset included also Russian articles, which were excluded from the subset used in our experiments.

interesting source articles.

Validation set of manually labeled positive and negative examples. For positive examples, we used the 21 manually identified interesting Latvian news 21P (see Section 3.3.1). However, no negative examples were provided. Therefore, for every Latvian article in the 21P collection we extracted five random articles, obtaining a list of 105 articles. The list was manually checked by a journalist from Ekspress Media who identified 38 articles as of no significance for retrieval. We denote these articles as **NL**. We combined the 21 Latvian examples from the 21P collection with the 38 negative articles from the **NL** set, which together form a validation set **V**.

We hypothesize that the articles of interest share common representation patterns. For every $k \in \{1, 100\}$ articles in `Latvianextracted@k` are used to learn a representation by using deep auto-encoder network architectures. We explore several deep auto-encoder network architectures. The main idea behind the network is that given the original representation of an article L_i the encoder part will encode the representation to a lower dimension, obtaining compressed intermediate representation C_{L_i} . The goal of the decoder is to learn to reconstruct the code back to an approximation of the original representation, L_i^* .

We consider using two different types of networks: regularized and non-regularized auto-encoder networks. We embed the articles with the Sentence-BERT (SBERT) (Reimers & Gurevych, 2019b) model (a modified pre-trained BERT (Devlin et al., 2019b) that uses a siamese and triplet network structure to derive semantically meaningful sentence embeddings that can be compared using cosine similarity) XLM-PASS-PHRASE in 768 dimensions, we use them as input. We use a 5-layer deep encoder architecture with dimensions $512 \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 32$. As for the decoder, we reverse the order of the same architecture. In all of the architectures, we use the RELU (Agarap, 2018) activation function between the layers. We optimize the $MeanSquaredError(L^*, L)$ as loss-function, with the Adam optimizer (Kingma & Ba, 2014), with learning rate of 0.001. We train for maximum of 1000 epochs, where we early-stop training if we do not improve the validation score in 10 consecutive epochs.

3.3.3 Classification settings

The auto-encoder outputs the reconstructions of the original input and cannot be used directly for classification. However, in many imbalanced classification (C. Zhang et al., 2016) and outlier detection (Chaurasia et al., 2020) problems, the auto-encoder is used to prioritize outputs based on its reconstruction error (via thresholding). We use the following scoring function:

$$g(L^*, L, t) = \begin{cases} 1 & \text{cosineSimilarity}(L^*, L) \geq t; \\ 0 & \text{otherwise;} \end{cases}$$

where L^* is the reconstructed and L the original representation, and classification threshold is denoted by t . In order to classify an example after a network is trained, we first reconstruct it through the network and then apply the classifying function g .

3.3.4 Threshold learning

In every learning epoch we first reconstruct the validation examples from the set V (21 positive and 37 negative gold standard examples) - obtaining reconstructed articles V^* . Next we measure the reconstruction errors and obtain a list of errors $R_{k,e}$, where k denotes the population size and e epoch. In order to decide on the classification threshold, we search the grid $\text{stepRange} = [\min(R_{k,e}), \max(R_{k,e})]$ with $\text{step} = 0.01$. We test every step value as a possible threshold value t . We first apply the classifying function g with t and measure the weighted F1-score of the classified reconstructions. We choose the t value such that we have the optimal F1-score. Formally, we choose t such that

$$\underset{t \in \text{stepRange}}{\operatorname{argmax}} \left[\text{F1-score} \left((g(V^*, V, t), \text{gold-standard}) \right) \right].$$

3.3.5 Evaluation

We evaluate the method in two scenarios, manual and automated. In both scenarios, we use the testing data for retrieving the top ranked articles as interesting and relevant.

Manual evaluation

We retrieve top-10 articles in two different network settings. The retrieved articles are manually evaluated by a journalist at Ekspress Meedia in the categories introduced in (Koloski, Zosa, et al., 2021), i.e. YES: the article is definitely relevant, MAYBE: the article is relevant to some extent and NO: the article is of no relevance. The results are described in Table 2.

The journalist found two articles as of definitive relevance (column “Yes”) and two of possible relevance (column “Maybe”) for retrieval in the best settings. Given that the problem is difficult, i.e. retrieving very special articles from a large set of all articles, the results still indicate that for Model32 (first line), 40% of the articles (four out of 10) are potentially interesting. This is slightly lower than the results of (Koloski, Zosa, et al., 2021), where in the best setting, one more article was labelled as MAYBE.

Table 2: Summary of the settings and the evaluations for the best-performing networks. The optimal threshold is shown in column “threshold”, followed by the number of epochs trained in the “epoch” column. Finally the F1 score represents the validation score, followed by the manual evaluations(YES/MAYBE/NO).

Name	Type	Train size	K-neigh	Threshold	Epoch	F1	Yes	Maybe	No
Model32	Non-regularized	712	10	0.6035	11	0.8093	2	2	6
Model32D	Regularized	1951	32	0.5961	5	0.7608	0	2	8
Baseline	majority-voting	x	x	x	x	0.4967	0	0	10

Automated evaluation

The goal of this experiment is to show that our method outscores random retrieval of articles. First, we construct a test-set consisted of the 21P labeled Latvian articles and the Latvian_{test} set. We randomly shuffle the articles in the new test-set. Next, we run the auto-encoder and measure the errors of reconstruction without applying threshold classification. Finally, we sort the articles by their reconstruction score in descending order. We search where the retrieved articles appear while obtaining k articles. We use Model32 to evaluate recall@k, where we treat the 21P articles as gold-standard. We also use random scoring of articles, to use it as a baseline. We execute 10⁶ random evaluations. The results in Figure 2 point out that our method outscores rankings obtained by retrieving random articles for translation. Implying that our method, works better than retrieving articles as interesting at random.

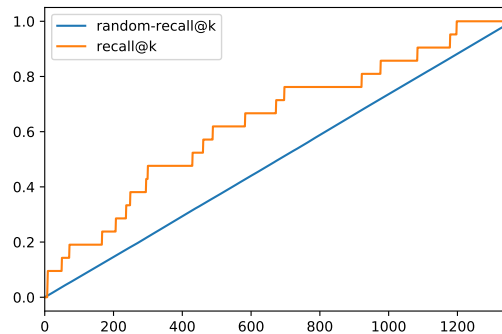


Figure 2: Automatic evaluation of the recall@k of the developed models for interesting cross-border news discovery retrieval. X-axis represents the number of documents evaluated, y-axis represents the cumulative recall@k score. The blue line represents the recall@k for random perturbation model. The orange line represents the recall@k for the chosen *Model32*. The *Model32* outcores the random perturbation model.

4 Final evaluation of EMBEDDIA tools on public datasets

4.1 Evaluation of autoBOT for the task of fake news classification against a generation of computer science students

We next discuss our recent evaluation of autoBOT (Škrlić et al., 2021), an AutoML (He et al., 2021; Gijbbers et al., 2019) system proposed as part of EMBEDDIA, against human competitors in a controlled environment. AutoML systems aim to, automatically, identify the best configuration of a given algorithm for a specified task. Even though commonly these systems aim to identify hyperparameter configurations, autoBOT aims to, alongside hyperparameter configurations, also identify sufficient representation combinations (there are many possible representations for a given document). The experiment was designed as follows. We considered the fake news classification data set proposed initially as a part of the CONSTRAINT workshop (Patwa et al., 2021). The data was first split in a stratified manner (70% – 30%). The system was benchmarked against the systems produced by the students of the third year of programs Computer Science and Computer Science and Mathematics at the Faculty of Computer and Information Science in Ljubljana; when they were given the data and the instructions to proceed with building the classifiers, they all witnessed approximately two full courses on artificial intelligence/machine learning methods, including natural language processing. The evaluation setting was as follows. A single train-test split was provided. The students were instructed that the classifiers may only be trained based on the training data, and finally tested on the test set. They had 20 days to produce the final solution (system capable of performing classification and its score on the test set). We measured the performance in accuracy, as the data was relatively balanced. We ran the default configuration of autoBOT for one hour on a workstation with 64 threads, used it to obtain the predictions on the test set (which we report as a strong baseline). The best model's final performance was 95.7%, other baselines given to the students can be seen in Table 3.

The final results were in favor of autoBOT. 54 groups of up to two students were not able to obtain Accuracy beyond 95.7%. Two groups of students, however, were able to obtain better models by fine-tuning large language models on Google's TPUs. One group considered ensembles of RoBERTa and BERT, and the other only BERT. Note that this group of better performing models (up to 97% accuracy for the ensembles) does not classify as “low-resource”, and requires specialized hardware.

Table 3: All baselines given to students for autoBOT vs. students evaluation.

Approach	Accuracy
autoBOT (1h) – the AutoML produced as part of EMBEDDIA	95.7%
MPNet + LR – Microsoft’s MPNet architecture-based representations used as input to a logistic regression classifier	93.9%
Char n-grams + LR – Character n-grams used as input to a logistic regression classifier	92.9%
Word n-grams + LR – Word n-grams used as input to a logistic regression classifier	91.2%
doc2vec + LR – doc2vec-based representations used as input to a logistic regression classifier	81.2%
Majority	52.3%

4.2 RuShiftEval: a shared task on semantic change detection for Russian

Words change their semantics over time as a result of combination of various processes that affect language simultaneously. Automatic detection and measuring the degree of meaning change could accelerate research in the history of language and also support a number of text analysis tasks connected to the EMBEDDIA project, such as media monitoring.

The SemEval Task on lexical semantic change detection (Schlechtweg et al., 2020), in which we participated with our own approach developed in the scope of the EMBEDDIA project (Martinc et al., 2020), provided valuable resources, i.e. datasets to compare various methods for semantic shift detection for four languages, English, German, Swedish, and Latin. However, results obtained on these datasets demonstrate high discrepancy: methods are ranked differently on different corpora and it is hard to find a single best-performing method.

UH then collaborated in organizing the next shared task for semantic change detection: RuShiftEval, a shared task on semantic change detection for Russian (Kutuzov & Pivovarova, 2021a). In the scope of the shared task, a novel gold standard dataset for diachronic viewpoint detection has been created (Kutuzov & Pivovarova, 2021b). The dataset consists of more than 100 Russian words manually annotated for time difference across three time periods: pre-soviet, soviet and post-soviet. As far as we are aware, this is the first semantic shift detection dataset which utilizes more than two time slices. This allows investigation on non-trivial nature of word meaning change: according to manual annotation, semantic shift between pre-soviet and post-soviet period cannot be calculated as a simple combination of change from pre-soviet to soviet and from soviet to post-soviet time periods.

The shared task was collocated with Dialogue 2021, the 27th International Conference on Computational Linguistics and Intellectual Technologies. The test and development datasets used in RuShiftEval are now publicly available, as well as the evaluation code and the baseline approach.

This work is described in full in (Kutuzov & Pivovarova, 2021a) and in (Kutuzov & Pivovarova, 2021b), attached here as Appendices B and C.

4.2.1 Dataset creation

For the competition, a new dataset of diachronic semantic changes for Russian words was created. Its novelty in comparison with prior work is its multiperiod nature. Until now, semantic change detection datasets focused on shifts occurring between two time periods. On the other hand, RuShiftEval provides human-annotated degrees of semantic change for a set of Russian nouns over three time periods: pre-Soviet (1700-1916), Soviet (1918-1990) and post-Soviet (1992-2016). Notably, it also contains ‘skipping’ comparisons of pre-Soviet meanings versus post-Soviet meanings. Together, this forms three subsets: RuShiftEval-1 (pre-Soviet VS Soviet), RuShiftEval-2 (Soviet VS post-Soviet) and RuShiftEval-3 (pre-Soviet VS post-Soviet). The three periods naturally stem from Russian history: they were radically

different in terms of life realities and writing and practices, which is reflected in the language. RuShiftEval can be used for testing the ability of semantic change detection systems to trace long-term multi-point dynamics of diachronic semantic shifts, rather than singular change values measured by comparing two time periods.

In building the dataset, we relied on the graded view on word meaning change (Schlechtweg et al., 2021): for each word in the dataset, we measure a degree of change between pairs of periods, rather than making a binary decision on whether its sense inventory changed over time. The measure relies on pairwise sentence annotations, where each pair of sentences is processed by at least three annotators.

Compiling the target-word set, we needed to ensure two main conditions: (i) the dataset contains many “interesting” words, i.e. words that changed their meaning between either pair of periods; (ii) not all words in the dataset actually changed their meaning. We followed the same procedure as in (Kutuzov & Kuzmenko, 2017; Rodina & Kutuzov, 2020; Schlechtweg et al., 2020): first, select changing words, and then augment them with fillers, i.e. random words following similar frequency distribution across three time periods. Technically, it would have been possible to populate the target word set automatically using any pre-trained language model (LM) for Russian and some measure of distance between word representations in different corpora. However, we wanted our target word choice to be motivated linguistically rather than influenced by a LM architecture.

The final dataset consists of 111 Russian nouns, where 12 words form a development set and 99 words serve as a test set. Annotators’ guidelines were identical to those in RuSemShift (Rodina & Kutuzov, 2020). To generate annotation tasks, we sampled 30 sentences from each sub-corpus and created sentence pairs containing two sentences from different time periods. We ran this sampling independently for all three period pairs. The sentences were accompanied by one preceding and one following sentence, to ease the annotators’ work in case of doubt. The task was formulated as labeling on a 1-4 scale, where 1 means the senses of the target word in two sentences are unrelated, 2 stands for ‘distantly related’, 3 stands for ‘closely related’, and 4 stands for ‘senses are identical’ (Hätyy et al., 2019). Annotators were also allowed to use the 0 (‘cannot decide’) judgments. They were excluded from the final datasets, but their number was negligible anyway: about 100 out of total 30 000.

The annotation was carried out on the Yandex.Toloka¹⁷ crowd-sourcing platform. We employed native speakers of Russian, older than 30, with a university degree. To ensure the annotation quality, the authors themselves annotated about 20 control examples for each pair of periods. We chose the most obvious cases of 1 and 4 for this. Annotators who answered incorrectly (not with the exactly matching grade) were banned from the task for 24 hours. The inter-annotator agreement (IAA) statistics and the number of judgments in each RuShiftEval subset are shown in Table 4.

Table 4: *RuShiftEval* statistics. α and ρ are inter-rater agreement scores as calculated by Krippendorff’s α (ordinal scale) and mean pairwise Spearman ρ . JUD is total number of judgments and 0-JUD is the number of 0-judgments (“cannot decide”).

Time bins	α	ρ	JUD	0-JUD
Test set (99 words)				
RuShiftEval-1	0.506	0.521	8 863	42
RuShiftEval-2	0.549	0.559	8 879	25
RuShiftEval-3	0.544	0.556	8 876	31
Development set (12 words)				
RuShiftEval-1	0.592	0.613	1 013	7
RuShiftEval-2	0.609	0.627	1 014	3
RuShiftEval-3	0.597	0.632	1 015	2

Each subset was annotated by about 100 human raters, more or less uniformly “spread” across annotation instances, with the only constraint being that each instance must be annotated by three differ-

¹⁷<https://toloka.yandex.ru/>

ent persons. Finally, the degrees of semantic change for each word between a pair of periods were calculated using the COMPARE metrics (Schlechtweg et al., 2018), which is the average of pairwise relatedness scores.

The dataset is publicly available, including the raw scores assigned by annotators¹⁸.

4.2.2 Task formulation for the shared task

In the RuShiftEval shared task the participants needed to rank a set of Russian words according to the strength of their meaning change, same as in Subtask 2 of the SemEval 2020 Task 1 (Schlechtweg et al., 2020). While in the past, this type of task has been tackled with unsupervised approaches (Schlechtweg et al., 2020; Rodina & Kutuzov, 2020), we encouraged participants to also consider developing a supervised approach towards solving the task, using the RuSemShift dataset (Rodina & Kutuzov, 2020) for training, in order to find out whether using training data actually helps semantic change detection. Submissions of the participants were processed, evaluated and ranked with the help of Codalab platform¹⁹.

During the main Evaluation phase (February 22 – March 1, 2021), the participants were provided with a set of 99 target Russian words to rank. During the Development phase (February 1 - February 22, 2021), a small development set was provided (12 manually annotated Russian words), and the participants could submit their predictions to get a preliminary estimation of their system performance (no gold labels were openly published). Before February 1, the shared task was in the Practice phase: the participants could submit predictions to the words from the RuSemShift test set (Rodina & Kutuzov, 2020). This dataset was already public, so the true labels were known to everyone. This phase could be used to sanity check submission routines.

Each participating team was able to submit up to 10 answers in the Evaluation phase, and up to 1000 answers in the Development phase. Submissions were evaluated using Spearman rank correlation between word ranking produced by a system and a gold ranking obtained in manual annotation. Thus, for each system we computed three correlations, for each of the time period pairs. The final ranking of the systems is based on averaging of the three scores.

4.2.3 Results

In the Evaluation phase, we received submissions from 14 users. Table 5 shows the performance of top submissions from each user or team (we give the name of the team by default or the name of the individual participant, if no team was associated with this submission). The teams are ranked by their average scores.

Among the best ranked teams, GlossReader (Rachinskiy & Arefyev, 2021) relied on the pretrained multilingual XLM-R language model (Conneau et al., 2019). On top of it, they trained a word sense disambiguation (WSD) system on English WSD datasets, using learned representations of sense definitions. Interestingly, this system shows excellent performance on Russian lexical semantic change data as well. Essentially, this participant reproduced the RuShiftEval annotation effort, replacing human judgments with the distances between XML-R contextualized embeddings of the target words. Additionally, a linear regression was trained on the RuSemShift dataset to convert vector distance values into relatedness scores (from 1 to 4).

DeepMistake (Arefyev et al., 2021) used the multilingual XLM-R as well, and also pre-trained on English WSD datasets, but without explicitly predicting senses. Similarly to GlossReader, they additionally fine-tuned this model on the RuSemShift using linear regression for mapping to relatedness scores.

¹⁸https://github.com/akutuzov/rushifteval_public

¹⁹<https://competitions.codalab.org/competitions/28340>

Table 5: *RuShiftEval* results (Spearman rank correlations). The Type column shows the type of the used distributional embeddings.

	Team	RuSemShift1	RuSemShift2	RuSemShift3	Mean	Type
1	GlossReader	0.781	0.803	0.822	0.802	token
2	DeepMistake	0.798	0.773	0.803	0.791	token
3	vanyatko	0.678	0.746	0.737	0.720	token
4	aryzhova	0.469	0.450	0.453	0.457	token
5	Discovery	0.455	0.410	0.494	0.453	token
6	UWB	0.362	0.354	0.533	0.417	type
7	dschlechtweg	0.419	0.373	0.383	0.392	type
8	jenskaiser	0.430	0.310	0.406	0.382	token
9	SBX-HY	0.388	0.281	0.439	0.369	type
	Baseline	0.314	0.302	0.381	0.332	type
10	svart	0.163	0.223	0.401	0.262	type
11	BykovDmitrii	0.274	0.202	0.307	0.261	token
12	fdzr	0.217	0.251	0.065	0.178	type

This is the first time that systems based on contextualized embeddings dominate the leaderboard. In both SemEval 2020 Task 1 (Schlechtweg et al., 2020) and DIACR-ITA (Basile et al., 2020), type embedding (or ‘static’ embedding) based architectures clearly won the rankings. In contrast, at RuShiftEval, five top performing systems use pre-trained contextualized (‘token-based’) models: XLM-R, BERT and ELMo. In the previous work, the researchers in the field expressed doubts about the abilities of token embeddings with relation to semantic change detection. It seems that at least in the case of RuShiftEval, they are able to solve the task better than their static counterparts. While this is encouraging in regards to the contextual embedding approach for semantic change detection that was developed in the scope of the Embeddia project (Montariol et al., 2021), the winning approaches also proposed several other improvements. The distinction between results in this shared task and results of the previous tasks most likely also lies in the difference between models rather than just between embeddings themselves.

Surprisingly, the first and the second best submissions relied on the contextualized XLM-R model (Conneau et al., 2019), which was not even specifically trained for processing Russian data. Its training corpus included texts in about 100 languages. Russian is well represented there but is far from being the largest in absolute size. The results of our shared task show that multilingual models like XLM-R can be very successfully applied to semantic change detection for Russian (and possibly for many other languages): their transferability is extremely high.

4.3 TREC 2021: News Track background linking task

With the massification of the internet and mobile devices, such as smartphones, people have started to access news more frequently from digital sources than printed ones (Stocking & Khuzam, 2021; Shearer, 2021). This has meant that newspaper publishers have had to focus more on the digital experience and perform users’ behavioral analysis for providing tools such as news recommendation (Wu et al., 2020). Furthermore, as Pranjić et al. (2020) indicate, linking news to other relevant articles can improve businesses’ websites metrics such as user engagement and average time on page. Subsequently, this can improve revenues from ads or sponsored articles.

Therefore, in 2018 the *Text REtrieval Conference (TREC)* along with *The Washington Post*²⁰, decided to propose the News Track (Soboroff et al., 2018), a track where the goal is to enhance users’ experience while reading news articles.

²⁰<https://www.washingtonpost.com/>

Since TREC 2020, the News Track is organized into two subtasks, *Background Linking* and *Wikification*. The former has been defined as the task where “given a news article, a system should retrieve other news articles that provide important context and/or background information that helps the reader better understand the query article” (Huang et al., 2018). The latter exploits, as a means of contextualization, the linking of textual elements, such as concepts and artifacts, to an external knowledge-base, in this case to Wikipedia (Soboroff et al., 2020).

In this section, we present our participation at the 2021 TREC News Track *Background Linking* task. Our participation consisted of five different approaches that used, for instance, keyword extraction, entities, and events detection, but also sentence embeddings and linear combination.

Our approaches are detailed in a paper by Cabrera-Diego et al. (Cabrera-Diego, Boros, & Doucet, 2021), attached to this deliverable as Appendix E.

4.3.1 Data

For 2021, the TREC News Track²¹ organizers provided a corpus composed of 728,626 news articles and blog posts published by *The Washington Post* from January 2012 through December 2020. Each document, either news article or blog post, includes elements such as title, kicker (section header), body, author, images captions, and publication date. Also, TREC organizers delivered a list of 51 different topics, i.e. news articles, for which TREC News Track’s participants had to propose background articles. For the 2021 edition of TREC News Track, the organizers also added a subtopic task, in which specific information, such as the background, is expected for each topic.

4.3.2 Data indexing

We first performed a pre-processing that consisted of parsing each document element, such as titles and captions, in order to get sentences. Once the documents were pre-processed, we decided to create embeddings for every document element using Sentence-BERT (Reimers & Gurevych, 2019c), a fine-tuned BERT (Devlin et al., 2019a) which produces embeddings that can be compared using cosine similarity. Specifically, we made use of the pre-trained model *stsb-mpnet-base-v2*²² which at the time of the experiments was the best performing model available. For this, we created composite vectors, in which we calculated the average embeddings based on multiple document elements: Title-Lead, Title-Body, and Title-Body-Captions. We also processed, in the same way, each topic provided by the TREC organizers, which notably included the creation of dense vectors for the narration or for the subtopics. For retrieving documents from the corpus, we indexed the pre-processed data using *Open Distro for Elasticsearch*²³ (ODFE), an *ElasticSearch*²⁴ branch which implements a k-NN algorithm that can be used to retrieve documents using dense vectors, such as embeddings.²⁵

4.3.3 Background linking approaches

For the Background Linking task, we proposed five different approaches described below.

Run 1: KWVec This approach consists of using keywords and dense vectors to retrieve the related background articles for a determined topic. Specifically, we start by extracting unigram keywords from

²¹<http://trec-news.org/>

²²<https://huggingface.co/sentence-transformers/stsb-mpnet-base-v2>

²³<https://opendistro.github.io>

²⁴<https://www.elastic.co/>

²⁵Although we use ODFE instead of ElasticSearch, the documentation of the latter is valid except for the dense vectors queries. Thus, we will point to ElasticSearch 7.12’s documentation in specific cases.

the text produced by the concatenation of the title, body, and captions.²⁶ This is done using YAKE (Campos et al., 2020), an unsupervised keyword extractor. Once we have the unigram keywords, we obtain those related to the title by matching the title's unigrams and the obtained keywords. The second step of KWVec consists of using a *boosting query*²⁷, where a collection of queries are used to retrieve the documents, and another set is used to decrease their relevance. To retrieve the documents, we submit three different queries to ODFE. Two of them ask ODFE to retrieve the documents that are relevant to the keywords found by YAKE. To be precise, we search title keywords in titles and body keywords in bodies. These queries are done through a *query string query*²⁸. Furthermore, as YAKE assigns a weight w to each keyword, we make use of these weights to increase or decrease the *query string query* relevance through the *boost* parameter. Nonetheless, as YAKE's weights interval is between $(0, \infty)$, where the lower the score the better, we modify it with Equation 1 to an interval of $(\inf, 0]$, where the higher the score the better.

$$KW_{weight} = \begin{cases} -\ln(w) & \text{if } w < 1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The third query retrieves the most relevant documents using ODFE's *exact k-NN* and cosine similarity²⁹. Specifically, the cosine similarity is calculated between the title-body dense vectors of the topic article and those found in the index. We modified ODFE's cosine similarity (s) score using Equation 2. The first reason is that ODFE's cosine similarity is vertically translated, within the interval $[0, 2]$, to provide only positive scores. The second reason is to boost the cosine similarity by a scalar defined experimentally to 250 and prevent its fading with respect to the keywords scores. More details are presented in Cabrera-Diego, Boros, & Doucet (2021).

$$Sim = \begin{cases} 250 \times (s - 1) & \text{if } s \geq 1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Run 2: Lambda Besides the previously described approach, we decided to explore a linear combination optimized through a Bayesian optimization algorithm (Moćkus et al., 1978)³⁰. Through this optimization, our goal was to determine the weights (λ) that different queries scores (x), such as title similarity, should be given in order to achieve the highest nDCG evaluation. This approach is similar to the one used by (Cabrera-Diego et al., 2014) for merging different systems outputs.

For the Lambda approach, we explored four different independent queries³¹, *title to title*, *body to body*, *lead to title* and *lead to body*, using two methods, keywords and dense vectors. This gave a total of eight different independent queries used for the optimization. The queries based on keywords use the method presented in Section 4.3.3, while queries based on dense vectors used an unmodified version of ODFE's *exact k-NN* and cosine similarity. To calculate the value of the different λ , we used as training data the sets provided by the organizers from previous years plus some additional articles that we annotated ourselves. The objective function to be maximized by the Bayesian optimization is presented in Equation 3, where G is a weighted harmonic average, Q_1 and Q_3 are respectively the first and third quartile, and Q_2 is the median. These values are calculated based on the nDCG@10 scores obtained

²⁶We concatenate these text fields in order to get more relevant keywords. Focusing separately on smaller text portions, such as the title, produced less relevant keywords.

²⁷<https://www.elastic.co/guide/en/elasticsearch/reference/7.12/query-dsl-boosting-query.html>

²⁸<https://www.elastic.co/guide/en/elasticsearch/reference/7.12/query-dsl-query-string-query.html>

²⁹<https://opendistro.github.io/for-elasticsearch-docs/docs/knn/knn-score-script/>

³⁰<https://github.com/fmfn/BayesianOptimization>

³¹This means that each query was done one by one.

by each topic for all the years (2018-2020).³²

$$G = \frac{5Q_1Q_2Q_3}{(Q_2Q_3) + (2.5Q_1Q_3) + (1.5Q_1Q_2)} \quad (3)$$

The weighted harmonic average presented in Equation 3 was defined to boost the median (Q_2) nDCG@10 score, but also to create a negatively skewed distribution of the nDCG@10 scores, by boosting the third quantile (Q_3). This would mean that we expect most of the nDCG@10 scores to have higher values rather than lower ones.

Run 3: 300K_ENT_PH This approach extends the KWVec method with a re-ranking step applied after the relevant documents were retrieved by the ODFE query. Thus, since named entity recognition (NER) has been playing an important role in information seeking and retrieval, we propose to exploit knowledge about entities and their relationships (events) for re-evaluating the relevance of the query results. For this and for taking advantage of the annotation efforts from previous campaigns, we leverage the fine-grained entities defined by the organizers of the TAC KBP *Recognizing Ultra Fine-grained Entities* (RUFES) 2020³³ and the events defined by the ACE 2005 evaluation campaign³⁴.

Fine-grained Entities. The KBP 2020 RUFES dataset provided by the organizers consisted of the development source documents and evaluation source documents drawn from a collection of The Washington Post news articles. The development source corpus and the evaluation source corpus had approximately 100,000 articles each, from which 50 documents were annotated for the development set with entity types from an ontology that contains approximately 200 fine-grained entity types and that followed the same three-level x.y.z hierarchy as in the TAC-KBP 2019 EDL track (Ji et al., 2019)³⁵. For example, such an entity organized in a hierarchy is: *Photographer* is from an *Artist* that, in turn, is a subtype of *PER*³⁶. In order to benefit from the extraction of these entity types, we made use of our recently proposed model for coarse-grained and fine-grained named entity recognition (Boros, Hamdi, et al., 2020; Boros, Pontes, et al., 2020; Boros, Hamdi, et al., 2021; Boros & Doucet, 2021) that consists in a hierarchical, multitask learning approach, with a fine-tuned encoder based on BERT (Devlin et al., 2019a).

Events. The annotated data of the ACE 2005 corpus provided by the ACE evaluation is restricted to a range of types, each with a set of subtypes. Thus, only the events of an appropriate type are annotated in a document. The eight event types (with 33 subtypes in parentheses) are: *Life* (*Be-Born, Marry, Divorce, Injure, Die*), *Movement* (*Transport*), *Conflict* (*Attack, Demonstrate*), etc. For detecting events, we focus on the event mention detection, and we use a BERT-based model with entity markers (Baldini Soares et al., 2019; Moreno et al., 2021, 2020; Boros, Moreno, & Doucet, 2021). This method is adapted from the BERT-based model with *EntityMarkers* (Baldini Soares et al., 2019) applied for relation classification, to perform event detection. The *EntityMarkers* model consists in augmenting the input data with a series of special tokens, e.g., if we consider a sentence $x = [x_0, x_1, \dots, x_n]$ with n tokens, we augment x with two reserved word pieces to mark the beginning and the end of each entity in the sentence. Thus, the previous sentence becomes: *There was free press in [GPE.Country_{start}] Qatar [GPE.Country_{end}], [ORG.CommercialOrganization_{start}] Al Jazeera [ORG.CommercialOrganization_{end}] but its' offices in [GPE.City_{start}] Kabul [GPE.City_{end}] and Baghdad were bombed by [ORG.Government.Agency_{start}] Americans [ORG.Government.Agency_{end}], where the different hierarchical entity types were detected by the previously presented model for fine-grained entity recognition.*

Re-ranking. For each sentence of the article, the entities and the event triggers are extracted and concatenated separated by a space, forming two separate text lines. Each line of entities or event triggers is encoded with Sentence-BERT and then, the final representation is the sum of all the obtained vectors

³²We explored different nDCG cuts, such as 50, 20 and 5. However, we found that, empirically, optimizing at 10 provided the best global results.

³³<https://tac.nist.gov/2020/KBP/RUFES/index.html>

³⁴<http://catalog.ldc.upenn.edu/LDC2006T06>

³⁵RUFES annotation guidelines: https://tac.nist.gov/2020/KBP/RUFES/guidelines/RUFES2020AnnotationGuidelines.v1.1_draft.pdf

³⁶*PER* refers to the entity type *Person*.

$v = (v_i)_{i=1}^n$ where each element $v_{i,j} = \sum_{j=1}^n x_{i,j}$. We use the cosine similarity for comparing the entity representations \cos .

Run 4: 300K_ENT_PH_DN This run is a re-ranking of the Run 3 (300K_ENT_PH) in which we include the cosine distances between the article text and the description and the narrative.

Run 5: Lambda_narr This run consisted in starting from the outcome produced by the Lambda approach and re-ranking the recommended articles using the narrative field. First, we calculated the cosine similarity between the narrative field dense vector and the recommended article's body dense vector. Then, we used a weighted harmonic mean to merge the rankings produced by the cosine similarity (R_{Narr}) and those produced by the Lambda approach (R_{Lambda}):

$$Lambda_narr = \frac{3.25R_{Lambda}^{-1}R_{Narr}^{-1}}{(2.25R_{Lambda}^{-1}) + R_{Narr}^{-1}} \quad (4)$$

We used the reciprocal of all the rankings R , to indicate that the lower the rankings, i.e. 1st, the better. In Equation 4 we give priority to the ranking produced by R_{Narr} over R_{Lambda} . To produce the final ranking, we sort $Lambda_narr$ scores in descending order.

4.3.4 TREC 2021 background linking results

In Table 6 we present, for each 2021 topic, the distribution of nDCG@5 scores calculated from all the submissions along with the scores obtained by each of our approaches while indicating the number of nDCG@5 scores, produced by our runs for each topic, found within each nDCG@5 quartiles. It should be noted, that in Table 6, if the value associated with a quartile was equal to another one, e.g. $Q_0 = Q_1$, like in topic 946, the score was assigned to the quartile closest to the median one (Q_2).

Table 6: Number of topics' nDCG@5 score found in each topic's quartile (Q) calculated by TREC organizers. The value in brackets represents the percentage of topics. Q_0 is the minimum score, Q_2 is the median and Q_4 is the maximum score.

Run	$x = Q_0$	$Q_0 < x < Q_1$	$Q_1 \leq x < Q_2$	$x = Q_2$	$Q_2 < x \leq Q_3$	$Q_3 < Q_4$	$x = Q_4$
KWVec	0 (0.0)	1 (1.9)	10 (19.6)	6 (11.7)	16 (31.3)	13 (25.4)	5 (9.8)
Lambda	1 (1.9)	3 (5.8)	12 (23.5)	4 (7.8)	11 (21.5)	13 (25.4)	7 (13.7)
300K_ENT_PH	0 (0.0)	0 (0.0)	8 (15.6)	5 (9.8)	17 (33.3)	16 (31.3)	5 (9.8)
300K_ENT_PH_DN	1 (1.9)	2 (3.9)	10 (19.6)	5 (9.8)	12 (23.5)	10 (19.6)	11 (21.5)
Lambda_narr	0 (0.0)	0 (0.0)	12 (23.5)	3 (5.8)	12 (23.5)	14 (27.4)	10 (19.6)

Based on the results present in Table 6, we can determine that the recommendations produced by our approaches generated an nDCG@5 greater than the participants' median in at least 60% of the topics. Specifically, KWVec 66.6%, Lambda 60.7%, 300K_ENT_PH 74.5%, 300K_ENT_PH_DN 64.7% and Lambda_narr 70.5%. Moreover, all our approaches achieved the maximum score nDCG@5 score in at least 9.8% of the topics, topped by 300K_ENT_PH_DN with a 21.5%. In regard to the re-rankings enhanced with entities and events or narratives, both runs, 300K_ENT_PH and 300K_PH_DN are rather homogeneous, with the same range of values [0.126, 0.714], and slightly similar median values. However, both Q_1 and Q_3 nDCG@5 scores surpass those of KWVec and Lambda. Despite the fact that model 300K_PH_DN achieved the largest number of topics with a maximum score, its median did not surpass that of KWVec's. In all the cases, the results obtained by 300K_ENT_PH and especially by 300K_PH_DN indicate that background linking can benefit from augmenting the articles with additional extracted information, such as named entities and events.

TREC 2021 News Track results All our methods had results that surpassed the best results in TREC 2021.

4.4 SemEval 2022 Task 8: Multilingual News Similarity

This SemEval task aims to develop systems that identify multilingual news articles that provide similar information. The task is: Given a pair of news articles (in the same language or in different languages), are they covering the same news story? This is a document-level similarity task in the applied domain of news articles, rating them pairwise on a real-valued [1-4] scale, from where 1 is most similar and 4 is least similar. We cover several techniques and propose different methods for finding the multilingual news article similarity by exploring the dataset in its entirety. We consider that the textual content, the provided metadata, and representative images corresponding to the news articles would draw on a multiplicity of modes, all of which contribute to the meaning and the main story of the news articles. Moreover, we also translate the articles in a high-resource language (English) in order to assess the ability of our models in an English-only context. Therefore, besides the articles, we took advantage of the article texts, the provided metadata (e.g., title, keywords, topics), the images (those that were available), and knowledge graph-based representations for entities and relations present in the articles. We investigate the multimodality of the data by experimenting with sentence, image, and knowledge graph embeddings by directly computing the semantic similarity between the different features and by predicting through regression the similarity scores.

4.4.1 Data

The training data has 4,964 article pairs from seven languages (English, German, Spanish, Arabic, Polish, Turkish, and French) and gold standard similarity scores for six dimensions (*Geography, Entities, Time, Narrative, Style, Tone*), plus the *Overall* score. The final evaluation data has 4,902 pairs and three “surprise” languages that were not present in the training data (Chinese, Italian, and Russian).

Table 7: Training and evaluation data statistics for SemEval 2022 Task 8.

	Train	Eval
Monolingual pairs	4,387	3,462
Cross-lingual pairs	577	1,440
Unseen language pairs	NA	2,000
Total	4,964	4,902
Top image	6,755	7,569

4.4.2 Approaches for assessing the similarity between news

We experiment with a variety of approaches for this task: document embeddings from Sentence-BERT (Reimers & Gurevych, 2019c) (pre-trained SBERT models and fine-tuned models) in both multilingual and monolingual settings, image embeddings, and knowledge graph embeddings. We evaluate the performance of the different models with the Pearson correlation between the similarity scores predicted by the model and the gold standard scores.

4.4.3 Semantic textual similarity models

A straightforward solution for finding the similarity between two texts is approaching it with sentence embeddings. Thus, we start our experimental setup by encoding the articles with Sentence-BERT (SBERT) (Reimers & Gurevych, 2019c). We explore this approach by encoding the articles with SBERT and using the cosine similarity of articles pairs as the predicted *Overall* score. For these experiments, we used the default hyperparameters provided by Reimers & Gurevych (2019c).

Table 8: Correlation between similarity scores from different proposed models and the *Overall* score.

Model		Pearson-r
<i>Semantic Textual Similarity & Regression</i>		
(1a) SBERT (PARAPHRASE-MULTILINGUAL-MPNET)	Similarity	0.6713
(1b) SBERT (ALL-MPNET) - Google Translate	Similarity	0.7139
(1c) SBERT (PARAPHRASE-MULTILINGUAL-MPNET)	Regression	0.7396
(1d) SBERT (ALL-MPNET) - Google Translate	Regression	0.7835
<i>Image Similarity & Regression</i>		
(2a) Images (CLIP-VIT-PATCH32)	Similarity	0.2991
(2b) Cross-images (CLIP-VIT-PATCH32)	Similarity	0.2607
(2c) Images (CLIP-VIT-PATCH32)	Regression	0.1043
(2d) Images (VIT-LARGE-PATCH32)	Regression	0.1124
<i>Knowledge Graph Similarity & Regression</i>		
(3a) KGm+LSA+SBERT (DISTILBERT+XLM-ROBERTA+ROBERTA)	Similarity	0.7128
(3b) KGm+LSA+SBERT (DISTILBERT)	Regression	0.5134
<i>Text & Image & Knowledge Graph Regression</i>		
(4a) Text+metadata (XLM-ROBERTA-LARGE)	Regression	0.7773
(4b) Text+metadata+images (XLM-ROBERTA-BASE+CLIP-VIT-PATCH32)	Regression	0.7020
(4c) Text+metadata+images (XLM-ROBERTA-LARGE+VIT-LARGE-PATCH32)	Regression	0.7335

Similarity based We first concatenate the title and the textual content of each article, and due to the multilingual characteristic of the data, we encode the textual sequence with a pre-trained multilingual SBERT model and compute the Pearson correlation between the cosine similarity of these sentence embeddings and the gold labels, results presented in Table 8 (1a). Then, we experiment with machine translating all the non-English articles to English using Google Translate and use an English SBERT model, results presented in Table 8 (1b).

Regression based We fine-tune the SBERT model on the multilingual pairs, results presented in Table 8 (1c) and on the machine-translated articles, results presented in Table 8 (1d). For fine-tuning, we use only the *Overall* score as the target similarity score. Since the similarity scores provided in the training data are in the range [1-4] from *most to least* similar, we normalize the *Overall* scores since the scores provided by cosine similarity are in the [0, 1] range from *least to most* similar.

Table 9: Extracts from an article pair where the Overall similarity score predicted by SBERT (1d) is 3.159, while the gold standard similarity score is 4.0. Similar terms in bold.

Article1	Article2
1492472369 (EN): At least one person has been confirmed dead , following Saturday's fire that gutted the Mgbuka Obosi Spare Parts Market in Idemili North Local Government Area of Anambra ... Mr Edwin Okadigbo , the Public Relations Officer on Tuesday ... Spokesperson of the Nigeria Security and Civil Defence Corps (NSCDC) , Anambra command, confirmed the incident in Awka.	1530831511 (EN): At least, one person has been confirmed dead ... in a road mishap that involved a commercial bus and a motorcycle in Mbosi junction, Ihiala Local Government Area of Anambra State ... Nigeria Security and Civil Defence Corps, NSCDC in Anambra State, Edwin Okadigbo said preliminary ...

Results We can substantially improve the English-only model for finding the similarity between two articles by finetuning not just with monolingual English pairs from the training data but by using all the machine-translated pairs. However, we observe some cases where our best performing finetuned model is misled by similar turns of phrase even if the article pair covers different events. We show extracts

from an article pair in Table 9 that covers a fire and a traffic accident, respectively. The gold *Overall* score for this pair is 4.0 (very dissimilar) but our best-performing model scores it at 3.1 (somewhat dissimilar) due to the similar phrasing that opens the articles and that they both mention the same-named entities.

4.4.4 Image similarity and regression models

We downloaded the images from the *top_image*, and as observed in Table 7, out of 9,928 articles (4,964 pairs), only 6,755 articles had a viable image in the train set, and out of 9,804 articles in the test set, only 7,567 were downloaded. For both, only around 60% of the articles had an image that could be used. Moreover, only around half of the pairs in both sets had representative images for both articles. Nevertheless, we attempted at using them in our approaches. We experiment with two recent pre-trained models, CLIP (Radford et al., 2021) and ViT (Dosovitskiy et al., 2020).

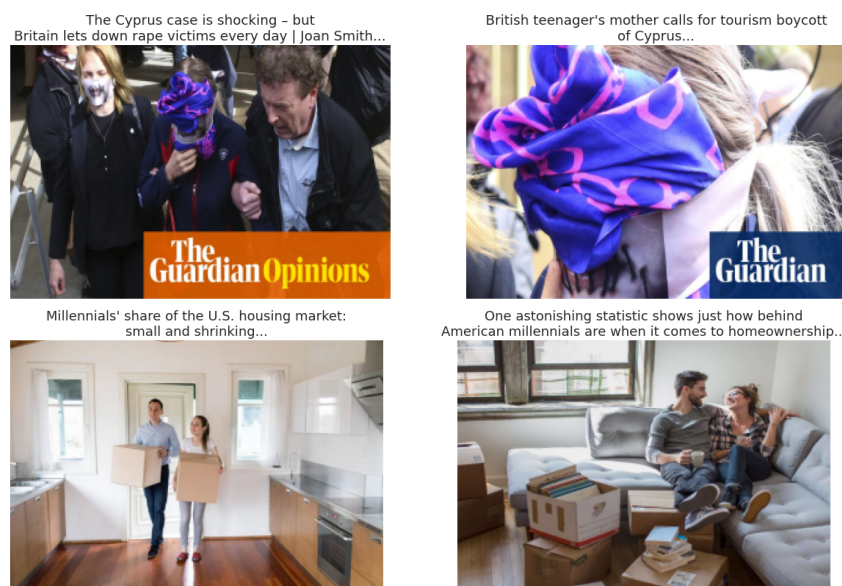


Figure 3: Two pairs of similar English articles (with a gold similarity score of 1.0 for both) correctly predicted by the image-based model (the pair from the upper figure has a similarity score of 1.28 & and the pair from the lower figure has a similarity score of 1.0). SBERT incorrectly predicted these similarity score (with a predicted similarity score of 1.83 for the upper figure and 1.63 for the lower figure).



Figure 4: A pair of marginally similar Russian articles (with a gold similarity score of 2.0), which is an unseen language during training, correctly predicted by the image-based model (with a predicted similarity score of 1.64), and incorrectly predicted by SBERT (with a predicted similarity score of 2.94).

Similarity based As for texts, we generate the image embeddings using CLIP, compute the cosine similarity between the paired images, and report the Pearson correlation between the obtained similarities and the gold labels. The results are presented in Table 8 (2a). As has been mentioned, many images are missed from the dataset. For those, we assign the default cosine similarity 0.5. However, we also tried an alternative strategy, which takes an advantage of the fact that CLIP is a multimodal model and produces images and text embeddings in the same space. Thus, we tried another strategy, called *Cross-images*. In this strategy, we compute all possible similarities between data points: image-to-image, text-to-text, and image-to-text. In the best case, when both images are available, this results in four similarities. In the worse case, only the similarity between texts is used. If only one image is available, the strategy results in two similarities: text-to-image and text-to-text. The final score is obtained by averaging the similarities available. Surprisingly, this strategy works slightly worse than an approach based solely on images, as can be seen in Table 8 (2b).

Regression based This method is detailed in Section 4.4.6. The results are presented in Table 8 (2c and 2d).

Results We analyzed the scores predicted by two textual-based methods, (1d) SBERT and (4a) Text+metadata, with the best scores when using only images (2a). Out of 4,902 pairs in the evaluation set (Table 7), only 2,009 had representative images for both news articles. Thus, we looked closer at the predictions for these pairs and noticed that 13% of them (262 pairs) were correctly predicted by the image-based model, and not by the text-based models, all of these being images with either faces or visible and clearly distinguished texts or text boxes, as shown in Figure 3 for two pairs of English articles. We also give an example where this model was able to better distinguish the similarity between two articles in an unseen language (Russian) in Figure 4, where the articles speak of the same topic but describe different events.

4.4.5 Knowledge graph similarity and regression models

We use the Wikidata5m (Wang et al., 2021) knowledge graph (KG) in order to retrieve knowledge-based features as used by Koloski, Stepišnik Perdih, et al. (2022). Similarly, we exploit six different knowledge graph-based embeddings: transE (Bordes et al., 2013), rotatE (Sun et al., 2019), complEx (Trouillon et al., 2016), distmult (Yang et al., 2015), simplE (Kazemi & Poole, 2018), and quate (S. Zhang et al., 2019). We use GraphVite (Zhu et al., 2019) pre-trained on aforementioned embeddings of the Wikidata knowledge graph. We concatenate the translated title and body of the articles to search n-grams of sizes 1, 2, and 3, as potential concepts appearing in the knowledge graph. After extracting potential candidates, we extract the embeddings of the candidates from the KG. In addition we generate latent semantic analysis (LSA), SBERT and stats representations as done by Koloski, Stepišnik-Perdih, et al. (2021). The results are in Table 8 (3a and 3b).

Similarity based First, we generate all ten feature spaces. Next, we generate combinations of feature spaces (1024 combinations in total), we concatenate and normalize them (KGm). Finally, we find thresholds to estimate the similarity scores, with respect to the *Overall* label. Our best results are presented in Table 8 (3a).

Regression based We utilize all six of the aforementioned KG representations, LSA and DistilBERT (Sanh et al., 2019) SBERT representations. In the next concatenation step, we have two different scenarios:

- Concatenation and normalisation of all of the inputs;
- Singular value decomposition (SVD) of the concatenated features to generate a new latent space of the devised features.

Next, we proceed to learn a deep neural network on the whole target space. Our best results are presented in Table 8 (3b).

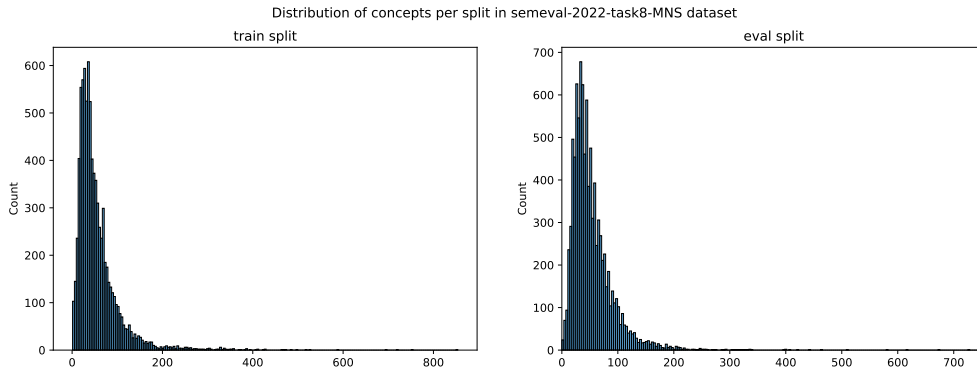


Figure 5: The distribution of concepts appearing in the train(left) and evaluation(right) split of the Wikidata5m Knowledge Graph. The x-axis counts the number of concepts per split, the y-axis counts the number of documents having that many concepts.

Results We analyzed the representations of articles based on the number of concepts retrieved per article and the top-most present articles. The top-most appearing concepts include entities such as *government, coronavirus, epidemic, report, information, death, economy, etc.*, showcasing us that most of the articles report about the pandemic, the statistics, and results. The distribution of concepts per document is shown in Figure 5. Originally, the Wikidata5m KG is based only on English concepts. Hence, we notice a performance drop in the representation of the translated articles in both of our approaches. In the future, we propose introducing multilingual KGs.

4.4.6 Text and image regression models

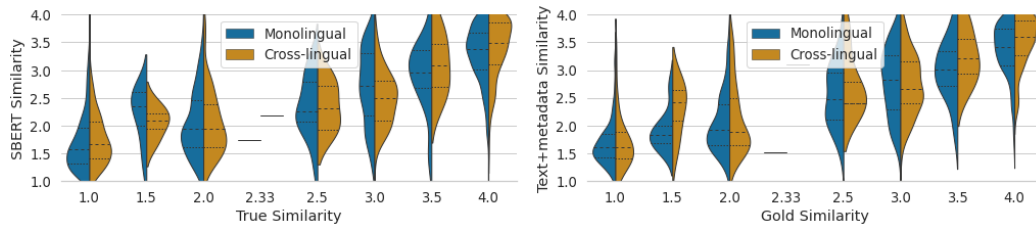


Figure 6: The distribution of gold and predicted similarity scores for the evaluation article pairs with available images for the text-based models, SBERT (1d from Table 8) and Text+metadata (4a from Table 8).

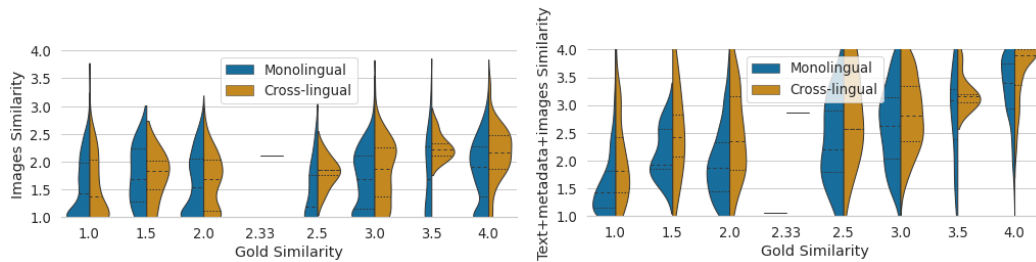


Figure 7: The distribution of gold and predicted similarity scores for the evaluation article pairs with available images for the image-based model, Images (3a from Table 8) and Text+metadata (4c from Table 8).

We also propose a classical approach that considers the task of finding the similarity between two articles by considering it as a regression task, and by predicting the similarity for all dimensions including *Overall*. This approach consists of a pre-trained and fine-tuned language model (Devlin et al., 2019b) and we encode the input in a particular manner. Because these models expect input data in a specific format, we need a special token, [SEP], to mark the end of a sentence or the separation between two sentences, and [CLS], at the beginning of a text generally used for classification or regression tasks.

Regression based After the pair of articles are tokenized and together encoded with [CLS] at the start and then separated by [SEP], they are passed through the encoder. Similarly, images are passed through a ViT encoder. For the missing images, we generate a *fake* white image. The BERT output token representations are afterward concatenated with the [CLS] representation and ViT output image representation and, fed to a linear layer for regression. The learning of the model is conducted end-to-end by optimizing an objective corresponding to *Overall* prediction. For these experiments, we utilized AdamW (Kingma & Ba, 2014) with a learning rate of 1×10^{-5} for two epochs with mean squared error (MSE) loss. We also considered a maximum sentence length of 512 (the maximum possible accepted by BERT or RoBERTa). These results are presented in Table 8 (from 4a to 4d).

Results Figures 6 and 7 present the Images (3a) similarity scores in comparison with Text+metadata (4b) and Text+metadata+images (4d) similarity scores. First, the results for Text+metadata (4a) seem to be rather similarly distributed to those provided by SBERT, with a slight difference the monolingual pairs with a gold score of 1.5, where SBERT generally predicts a similarity of 2.5. When using image representations, not surprisingly, we notice that the results for Images (2a) are generally staying around and average of 2.0, proving that having only around half of the train and test sets with images is not enough in helping distinguishing news articles. We also observe that Text+metadata+images scores are influenced by the lack or by the presence of images, improving the results for the most dissimilar pairs, with 4.0 and 3.5 scores, while skewing the distribution for both monolingual and cross-lingual pairs with a gold similarity of 2.5 and 2.0 (this is probably due to the fact that not having images for almost half of articles brought the scores towards an average similarity).

Ranking at SemEval-2022 Task 8 In the official SemEval-2022 Task-8, we ranked fifth in the overall team ranking multilingual and cross-lingual results, and second in the English-only results, both with our *Semantic Textual Similarity* with pre-trained multilingual and monolingual SBERT models.

4.5 Multilingual Topic Labelling of News Topics using Ontological Mapping

Topic models uncover the latent themes in a document collection through the co-occurrences of words in documents. The large volume of news produced daily makes topic modelling useful for analysing topical news 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 assign labels to topics to improve interpretability.

Topic labelling is the task of assigning a short label to a topic that captures its semantic content. Various methods have been proposed to generate topic labels but there has been no work so far on creating 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 (i.e. IPTC NewsCodes). 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 without modifications to topics from languages that have not been seen during training.

Full details are given in (Zosa et al., 2022), attached to this deliverable as Appendix D.

4.5.1 Models

Ontology mapping (our approach). 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 multi-label classification task where a topic can be classified as belonging to one or more concepts in the ontology.

We encode the top terms of the topic using SBERT (Reimers & Gurevych, 2019a) and pass this representation to a neural network classifier. We refer to this as the **ontology** model.

Comparisons to state-of-the-art. We investigate how our ontology mapping method compares to methods that directly generate topic labels. We implement a RNN seq2seq model using the same hyperparameters as (Alokaili et al., 2020). We refer to this as the **rnn** model. We also implement a slightly modified model where we replace RNN with transformers. We refer to this as the **transformer** model.

We also finetune mBART (Liu et al., 2020), and set the source and target languages to Finnish. We finetuned mBART-25 from HuggingFace³⁷ for 5 epochs. We refer to this as the **mbart** model.

4.5.2 Datasets

News Ontology. We use the IPTC NewsCodes as our news ontology.³⁸ 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. We use news articles from 2017 of the Finnish News Agency (STT) dataset (STT, 2019; STT et al., 2020) which have been tagged with IPTC concepts. We construct a dataset where the top n words of an article are treated as input $X = (x_1, \dots, x_n)$ 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).

Test Data and Gold labels. 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 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. This dataset of Finnish news topics and gold standard labels will be available in the Github repository associated with this work.³⁹ Our annotators are compensated for their efforts. We limited our test data to 30 topics due to budget constraints.

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 (Bhatia et al., 2016). 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.

4.5.3 Results

We use BERTScore (T. Zhang et al., 2019) to evaluate the labels generated by the models with regards to the gold standard labels. BERTScore finds optimal correspondences between gold standard tokens

³⁷<https://huggingface.co/facebook/mbart-large-cc25>

³⁸<https://cv.iptc.org/newscodes/subjectcode/>

³⁹<https://github.com/ezosa/topic-labelling>

Table 10: Averaged BERTScores between labels generated by the models and the gold standard labels for Finnish and English news topics.

	P	R	F1
Finnish news			
<i>baseline: top 5 terms</i>	89.47	88.08	88.49
ontology-tfidf (ours)	94.54	95.42	94.95
ontology-sent (ours)	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 (zero-shot)			
<i>baseline: top 5 terms</i>	98.17	96.58	97.32
ontology-tfidf	97.00	95.25	96.04
ontology-sent	97.18	95.43	96.21

Table 11: Generated labels for a topic in Finnish news. Finnish labels are manually translated except for ontology-sent. For ontology-sent, we provide the IPTC concept ID and the corresponding Finnish and English labels.

	Finnish topic
Top topic words	räikkönen, bottas, ajaa (<i>to drive</i>), hamilton, mercedes
Gold	formula, formulat, formula 1, f1, formula-auto, aika-ajot (<i>time trial</i>), moottoriurheilu (<i>motor sport</i>)
rnn-tfidf	autourheilu (<i>auto sport</i>), urheilutapahtumat (<i>sports event</i>), mm-kisat (<i>world championship</i>), urheilu (<i>sport</i>), urheilijat (<i>athletes</i>)
transformer-sent	urheilutapahtumat (<i>sports event</i>), mm-kisat (<i>world championship</i>), urheilu (<i>sport</i>), autourheilu (<i>auto sport</i>), kansainväliset (<i>international</i>)
mbart-sent	autourheilu moottoriurheilu, urheilutapahtumat, mm-kisat , urheilijat pelaajat, urheilu
ontology-sent (ours)	ID: 15000000, fi: <u>urheilu</u> , en: sport; ID: 15039000, fi: <u>autourheilu moottoriurheilu</u> , en: motor racing; ID: 15073000, fi: <u>urheilutapahtumat</u> , en: sports event; ID: 15039001, fi: <u>formula 1</u> , en: formula one; ID: 15073026, fi: <u>mm-kisat</u> , en: world championship

and generated tokens and from these correspondences, precision (P), recall (R), and F1 scores are computed.

We show the BERTScores for the Finnish news topics at the top of Table 10. All models outperform the baseline by a large margin. This shows that ontology concepts are, as labels, better aligned with human-preferred labels than the top topic words. We do not see a significant difference in performance between training on the tfidf or sent datasets.

In Table 11 (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.

BERTScore results for English topics are shown at the bottom of Table 10. 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.

4.5.4 Conclusions

We proposed 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 the IPTC NewsCodes, 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 show that labels of ontology concepts correlate highly with labels preferred by humans. We also release a novel dataset of Finnish news topics with gold standard labels.

5 Conclusions

This deliverable presented our media partners' datasets published in WP4 and evaluations performed thereon.

Published articles are a valuable resource for further scientific research contributing to low-resource languages such as Estonian, Croatian and Finnish. We also enriched a subset of our public dataset with the results of our developed tools in order to make the results available, easily analysable and usable.

Our media partners in Estonia (Ekspress Meedia) evaluated manually the results of our two keyword extraction methods. TNT-KID received higher scores than RaKUn, which is in line with the expectations of supervised methods performing better than the unsupervised ones. TNT-KID results are sufficiently good to be integrated in the life production. RaKUn on the other hand, is already used in a National Library of Estonia tender for automatic subject indexing (this means keyword tagging for books, brochures, articles etc). On our media partners' articles we also performed an evaluation of sentiment analysis that performed substantially better than simple majority-class classifier and cross-border news extraction that outscored rankings obtained by retrieving random articles for translation.

In order to further evaluate the tools for semantic change detection, we organized a shared task RuShiftEval, where the participants had to rank a set of Russian words according to their diachronic change. Interestingly, this was the first shared task on the topic of semantic change, in which the systems employing contextualized embeddings overwhelmingly outranked systems that employed static embeddings. Another interesting finding, especially important for the EMBEDDIA project, is that participants of the shared tasks that applied multilingual models, ranked first and second.

From our participation to TREC 2021, we noticed that, despite the existence of embeddings from fine-tuned language models such as Sentence-BERT (Reimers & Gurevych, 2019c), keywords are still one of the most powerful sources of knowledge to rank news articles.

This report also contains an evaluation of the AutoML tool autoBOT against human competitors – the results of the internal competition where the autoBOT's performance was compared to a collection of solutions proposed by upper-undergraduate students indicate competitive performance with minimal human (developer) input. We believe autoBOT enables more widespread use of state-of-the-art representation learning techniques for data scientists that are not necessarily experienced in the domain of NLP.

In our investigation of methods for evaluating the similarity of multilingual news articles and our participation to SemEval 2022, we found that finetuning large pretrained Transformer-based models (e.g. Sentence-BERT) performed best. Using news metadata such as images also yielded promising results although these kinds of data might not always be available for all news articles.

We also developed a novel method for producing labels for news topics in multiple languages, to help exploring multilingual news content. We also produced a new Finnish dataset of news topics and gold standard labels based on the STT news collection.

6 Associated Outputs

The work described in this deliverable has resulted in the following resources:

Description	URL	Licence
Cross-border news discovery	https://github.com/bkolosk1/Interesting-cross-border-news-discovery	MIT
Multilingual topic labelling	https://github.com/ezosa/topic-labelling	MIT
Ekspress Meedia News Archive (in Estonian and Russian) 1.0	http://hdl.handle.net/11356/1408	CC BY-NC-ND 4.0
Latvian Delfi article archive (in Latvian and Russian) 1.0	http://hdl.handle.net/11356/1409	CC BY-NC-ND 4.0
24sata news article archive 1.0	http://hdl.handle.net/11356/1410	CC BY-NC-ND 4.0
Finnish News Agency Archive 1992-2018	http://urn.fi/urn:nbn:fi:lb-2019041501	CLARIN RES end-user license +NC +OTHER 2.0
Finnish News Agency Archive 1992-2018, CoNLL-U	http://urn.fi/urn:nbn:fi:lb-2020031201	CLARIN RES end-user license +NC +OTHER 2.0
Finnish News Agency Archive 2019-2021	https://metashare.csc.fi/repository/browse/finnish-news-agency-archive-2019-2021-source/ee6145c2882211eca1f5fa163ec5ae3e1d0fa3d38e314897b1	CLARIN RES end-user license +NC +OTHER 2.0
Keyword extraction datasets for Croatian, Estonian, Latvian and Russian 1.0	http://hdl.handle.net/11356/1403	CC BY-NC-ND 4.0
Sentiment Annotated Dataset of Croatian News	http://hdl.handle.net/11356/1342	CC BY-NC-ND 4.0
Estonian-Latvian Interesting News Pairs	https://github.com/EMBEDDIA/interesting-cross-border-news-discovery	MIT
Computer-Generated Statistical News Texts	https://github.com/EMBEDDIA/embeddia-nlg-output-corpus	Attribution-ShareAlike 4.0 International
EMBEDDIA tools output example corpus of Estonian, Croatian and Latvian news articles 1.0	http://hdl.handle.net/11356/1485	CC BY-NC-ND 4.0
SemEval 2022 Task 8: Multilingual news similarity	https://github.com/bkolosk1/semeval-2022-MNS	MIT

The work described in this deliverable has resulted also in the following publications, added in this deliverable as appendices:

Citation	Status	Appendix
Pollak, Senja et al. (2021). EMBEDDIA Tools, Datasets and Challenges: Resources and Hackathon Contributions. In <i>Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation</i> (pp. 99-109).	Published	Appendix A
Kutuzov, Andrey & Pivovarov, Lidia (2021). RuShiftEval: a shared task on semantic shift detection for Russian. <i>Computational linguistics and intellectual technologies:: Papers from the annual conference Dialogue</i> .	Published	Appendix B
Kutuzov, Andrey & Pivovarov, Lidia (2021). Three-part diachronic semantic change dataset for Russian. In <i>Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021</i> (pp. 7-13).	Published	Appendix C
Zosa, Elaine and Pivovarov, Lidia and Boggia, Michele & Ivanova, Sardana (2022). Multilingual Topic Labelling of News Topics using Ontological Mapping. In <i>Proceedings of the 44th European Conference on Information Retrieval</i>	To appear	Appendix D
Cabrera-Diego, Luis Adrián, & Boros, Emanuela & Doucet, Antoine (2022, February). Elastic Embedded Background Linking for News Articles with Keywords, Entities and Events. In <i>Proceedings of the Text Retrieval Conference (TREC) 2021</i>	To appear	Appendix E

References

- Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.
- Alokaili, A., Aletras, N., & Stevenson, M. (2020). Automatic generation of topic labels. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1965–1968).
- Arefyev, N., Fedoseev, M., & Protasov, V. (2021). Deepmistake: Which senses are hard to distinguish for a word-in-context model. *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialogue conference*.
- Baldini Soares, L., FitzGerald, N., Ling, J., & Kwiatkowski, T. (2019, July). Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2895–2905). Florence, Italy: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/P19-1279> doi: 10.18653/v1/P19-1279
- Basile, V., Di Maro, M., Danilo, C., & Passaro, L. C. (2020). Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for Italian. In *Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian* (pp. 1–7).
- Bhatia, S., Lau, J. H., & Baldwin, T. (2016). Automatic labelling of topics with neural embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 953–963).
- Bordes, A., Usunier, N., García-Durán, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States* (pp. 2787–2795). Retrieved from <https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html>
- Boros, E., & Doucet, A. (2021). Transformer-based methods for recognizing ultra fine-grained entities (rufes). *arXiv preprint arXiv:2104.06048*.
- Boros, E., Hamdi, A., Linhares Pontes, E., Cabrera-Diego, L. A., Moreno, J. G., Sidere, N., & Doucet, A. (2020, November). Alleviating Digitization Errors in Named Entity Recognition for Historical Documents. In Raquel Fernández & Tal Linzen (Eds.), *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL)* (pp. 431–441). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.35
- Boros, E., Hamdi, A., Pontes, E. L., Cabrera-Diego, L. A., Moreno, J., Sidere, N., & Doucet, A. (2021). Atténuer les erreurs de numérisation dans la reconnaissance d'entités nommées pour les documents historiques. In *CONFérence en Recherche d'Informations et Applications-CORIA 2021, French Information Retrieval Conference*. Online: CORIA.
- Boros, E., Moreno, J. G., & Doucet, A. (2021). Event detection with entity markers. In D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, & F. Sebastiani (Eds.), *Advances in Information Retrieval* (pp. 233–240). Cham: Springer International Publishing.

- Boros, E., Pontes, E. L., Cabrera-Diego, L. A., Hamdi, A., Moreno, J., Sidère, N., & Doucet, A. (2020). Robust named entity recognition and linking on historical multilingual documents. In L. Cappellato, C. Eickhoff, N. Ferro, & A. Névél (Eds.), *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum* (Vol. 2696, pp. 1–17). Thessaloniki, Greece: CEUR-WS.
- Cabrera-Diego, L. A., Boros, E., & Doucet, A. (2021). Elastic embedded background linking for news articles with keywords, entities and events. In *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC* (pp. 16–20).
- Cabrera-Diego, L. A., Huet, S., Jabaian, B., Molina, A., Torres-Moreno, J.-M., El-Bèze, M., & Durette, B. (2014). Algorithmes de classification et d'optimisation : participation du LIA/ADOC à DEFT'14. In *Actes du dixième Défi Fouille de Textes* (pp. 53–60). Marseille, France.
- Cabrera-Diego, L. A., Moreno, J. G., & Doucet, A. (2021, April). Using a frustratingly easy domain and tagset adaptation for creating Slavic named entity recognition systems. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing* (pp. 98–104). Kiyv, Ukraine: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.bsnlp-1.12>
- Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509, 257–289. doi: 10.1016/j.ins.2019.09.013
- Chaurasia, S., Goyal, S., & Rajput, M. (2020). Outlier detection using autoencoder ensembles: A robust unsupervised approach. In *2020 International Conference on Contemporary Computing and Applications (IC3A)* (p. 76-80). doi: 10.1109/IC3A48958.2020.233273
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019a). BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. Retrieved from <https://doi.org/10.18653/v1/n19-1423> doi: 10.18653/v1/n19-1423
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019b, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/N19-1423> doi: 10.18653/v1/N19-1423
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Freienthal, L., Pelicon, A., Martinc, M., Škrlić, B., Krustok, I., Pranjić, M., ... Koloski, B. (2022). *EMBED-DIA tools output example corpus of Estonian, Croatian and Latvian news articles 1.0*. Retrieved from <http://hdl.handle.net/11356/1485>
- Gijssbers, P., LeDell, E., Thomas, J., Poirier, S., Bischl, B., & Vanschoren, J. (2019). An open source automl benchmark. *arXiv preprint arXiv:1907.00909*.
- Hätty, A., Schlechtweg, D., & Schulte im Walde, S. (2019, June). SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)* (pp. 1–8). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/S19-1001> doi: 10.18653/v1/S19-1001

- He, X., Zhao, K., & Chu, X. (2021). Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212, 106622.
- Huang, S., Soboroff, I., & Harman, D. (2018). TREC 2018 News Track. In Dyaa Albakour, David Corney, Julio Gonzalo, Miguel Martinez, Barbara Poblete, & Andreas Valochas (Eds.), *Proceedings of the Second International Workshop on Recent Trends in News Information Retrieval (NewsIR'18)* (Vol. 2079). Grenoble, France: CEUR Workshop Proceedings. Retrieved from <http://ceur-ws.org/Vol-2079/paper12.pdf>
- Ji, H., Sil, A., Dang, H. T., Soboroff, I., Nothman, J., & Hub, S. I. (2019). Overview of tac-kbp2019 fine-grained entity extraction. In *2019 Text Analysis Conference Proceedings*. Gaithersburg, Maryland, USA: NIST.
- Kazemi, S. M., & Poole, D. (2018). Simple embedding for link prediction in knowledge graphs. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada* (pp. 4289–4300). Retrieved from <https://proceedings.neurips.cc/paper/2018/hash/b2ab001909a8a6f04b51920306046ce5-Abstract.html>
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Koloski, B., Pollak, S., Škrlić, B., & Martinc, M. (2021a, April). Extending neural keyword extraction with TF-IDF tagset matching. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation* (pp. 22–29). Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.hackashop-1.4>
- Koloski, B., Pollak, S., Škrlić, B., & Martinc, M. (2021b). *Keyword extraction datasets for Croatian, Estonian, Latvian and Russian 1.0*. Retrieved from <http://hdl.handle.net/11356/1403> (Slovenian language resource repository CLARIN.SI)
- Koloski, B., Pollak, S., Škrlić, B., & Martinc, M. (2022). *Out of Thin Air: Is Zero-Shot Cross-Lingual Keyword Detection Better Than Unsupervised?*
- Koloski, B., Stepišnik-Perdih, T., Pollak, S., & Škrlić, B. (2021). Identification of covid-19 related fake news via neural stacking. In T. Chakraborty, K. Shu, H. R. Bernard, H. Liu, & M. S. Akhtar (Eds.), *Combating Online Hostile Posts in Regional Languages during Emergency Situation* (pp. 177–188). Cham: Springer International Publishing.
- Koloski, B., Stepišnik-Perdih, T., Robnik-Šikonja, M., Pollak, S., & Škrlić, B. (2022). Knowledge graph informed fake news classification via heterogeneous representation ensembles. *Neurocomputing*. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0925231222001199> doi: <https://doi.org/10.1016/j.neucom.2022.01.096>
- Koloski, B., Zosa, E., Stepišnik-Perdih, T., Škrlić, B., Paju, T., & Pollak, S. (2021). Interesting cross-border news discovery using cross-lingual article linking and document similarity. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation* (pp. 116–120).
- Kutuzov, A., & Kuzmenko, E. (2017). Two centuries in two thousand words: neural embedding models in detecting diachronic lexical changes. In *Quantitative approaches to the russian language* (pp. 95–112). Routledge.
- Kutuzov, A., & Pivovarova, L. (2021a). RuShiftEval: a shared task on semantic shift detection for Russian. In *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*.
- Kutuzov, A., & Pivovarova, L. (2021b). Three-part diachronic semantic change dataset for russian. *arXiv preprint arXiv:2106.08294*.

- Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., ... Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8, 726–742.
- Martinc, M., Montariol, S., Zosa, E., Pivovarova, L., et al. (2020). Discovery team at semeval-2020 task 1: Context-sensitive embeddings not always better than static for semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*.
- Martinc, M., Škrlić, B., & Pollak, S. (2021). Tnt-kid: Transformer-based neural tagger for keyword identification. *Natural Language Engineering*, 1–40. doi: 10.1017/S1351324921000127
- Montariol, S., Martinc, M., & Pivovarova, L. (2021). Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4642–4652).
- Moreno, J. G., Boros, E., & Doucet, A. (2020). Tlr at the ntcir-15 finnum-2 task: Improving text classifiers for numeral attachment in financial social data. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies* (pp. 8–11). Tokyo, Japan: National Institute of Informatics.
- Moreno, J. G., Doucet, A., & Grau, B. (2021, January). Relation classification via relation validation. In *Proceedings of the 6th Workshop on Semantic Deep Learning (SemDeep-6)* (pp. 20–27). Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.semdeep-1.4>
- Močkus, J., Tiešis, V., & Žilinskas, A. (1978). The application of Bayesian methods for seeking the extremum. In G. P. Szegő & L. C. W. Dixon (Eds.), *Towards Global Optimisation* (Vol. 2, pp. 117–128). North-Holland.
- Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., ... Chakraborty, T. (2021). Fighting an infodemic: Covid-19 fake news dataset. *Communications in Computer and Information Science*, 21–29. Retrieved from http://dx.doi.org/10.1007/978-3-030-73696-5_3 doi: 10.1007/978-3-030-73696-5_3
- Pelicon, A., Pranjić, M., Miljković, D., Škrlić, B., & Pollak, S. (2020b). *Sentiment Annotated Dataset of Croatian News*. Retrieved from <http://hdl.handle.net/11356/1342> (Slovenian language resource repository CLARIN.SI)
- Pelicon, A., Pranjić, M., Miljković, D., Škrlić, B., & Pollak, S. (2020a, Aug). Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, 10(17), 5993. Retrieved from <http://dx.doi.org/10.3390/app10175993> doi: 10.3390/app10175993
- Pollak, S., Purver, M., Shekhar, R., Freienthal, L., Kuulmets, H.-A., & Krustok, I. (2021). *Latvian Delfi article archive (in Latvian and Russian) 1.0*. Retrieved from <http://hdl.handle.net/11356/1409> (Slovenian language resource repository CLARIN.SI)
- Pollak, S., Robnik-Šikonja, M., Purver, M., Boggia, M., Shekhar, R., Pranjić, M., ... others (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* (pp. 99–109).
- Pranjić, M., Podpečan, V., Robnik-Šikonja, M., & Pollak, S. (2020, September). Evaluation of related news recommendations using document similarity methods. In D. Fišer & T. Erjavec (Eds.), *Proceedings of the Conference on Language Technologies and Digital Humanities (JDTH2020)* (pp. 81–86). Ljubljana, Slovenia: Inštitut za novejšo zgodovino. doi: 10.5281/zenodo.4059710
- Purver, M., Pollak, S., Freienthal, L., Kuulmets, H.-A., Krustok, I., & Shekhar, R. (2021). *Ekspress news article archive (in Estonian and Russian) 1.0*. Retrieved from <http://hdl.handle.net/11356/1408> (Slovenian language resource repository CLARIN.SI)

- Purver, M., Shekhar, R., Pranjić, M., Pollak, S., & Martinc, M. (2021). *24sata news article archive 1.0*. Retrieved from <http://hdl.handle.net/11356/1410> (Slovenian language resource repository CLARIN.SI)
- Rachinskiy, M., & Arefyev, N. (2021). Zeroshot crosslingual transfer of a gloss language model for semantic change detection. *Komp'yuternaya Lingvistika i Intellekturnye Tekhnologii: Dialogue conference*.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748–8763).
- Reimers, N., & Gurevych, I. (2019a). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992).
- Reimers, N., & Gurevych, I. (2019b, 11). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. Retrieved from <https://arxiv.org/abs/1908.10084>
- Reimers, N., & Gurevych, I. (2019c, November). 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). Hong Kong, China: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/D19-1410> doi: 10.18653/v1/D19-1410
- Rodina, J., & Kutuzov, A. (2020, December). RuSemShift: a dataset of historical lexical semantic change in Russian. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1037–1047). Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved from <https://aclanthology.org/2020.coling-main.90> doi: 10.18653/v1/2020.coling-main.90
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020, December). SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 1–23). Barcelona (online): International Committee for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/2020.semeval-1.1>
- Schlechtweg, D., Schulte im Walde, S., & Eckmann, S. (2018, June). Diachronic usage relatedness (DUREl): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 169–174). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/N18-2027> doi: 10.18653/v1/N18-2027
- Schlechtweg, D., Tahmasebi, N., Hengchen, S., Dubossarsky, H., & McGillivray, B. (2021). Dwug: A large resource of diachronic word usage graphs in four languages. *arXiv preprint arXiv:2104.08540*.
- Shearer, E. (2021, December). More than eight-in-ten Americans get news from digital devices. *Pew Research Center*. Retrieved 2021-09-24, from <https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/>
- Škrlić, B., Martinc, M., Lavrač, N., & Pollak, S. (2021). autobot: evolving neuro-symbolic representations for explainable low resource text classification. *Machine Learning*, 110(5), 989–1028.

- Soboroff, I., Huang, S., & Harman, D. (2018). TREC 2018 News Track Overview. In E. M. Voorhees & A. Ellis (Eds.), *Proceedings of the 27th Text REtrieval Conference (TREC 2018)* (Vol. SP 500-331). Gaithersburg, Maryland, USA: NIST. Retrieved from <https://trec.nist.gov/pubs/trec27/papers/Overview-News.pdf>
- Soboroff, I., Huang, S., & Harman, D. (2020). TREC 2020 News Track Overview. In E. M. Voorhees & A. Ellis (Eds.), *Proceedings of the 29th Text REtrieval Conference (TREC 2020)* (Vol. SP 1266). Online: NIST. Retrieved from <https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.N.pdf>
- Stocking, G., & Khuzam, M. (2021, June). Digital News Fact Sheet. *Pew Research Center*. Retrieved 2021-09-24, from <https://www.pewresearch.org/journalism/fact-sheet/digital-news/>
- STT. (2019). *Finnish News Agency Archive 1992-2018, source* (<http://urn.fi/urn:nbn:fi:lb-2019041501>). Kielipankki.
- STT. (2022). *Finnish News Agency Archive 2019-2021*. Kielipankki.
- STT, Helsingin yliopisto, & Alnajjar, K. (2020). *Finnish News Agency Archive 1992-2018, CoNLL-U, source* (<http://urn.fi/urn:nbn:fi:lb-2020031201>) [text corpus]. Kielipankki. Retrieved from <http://urn.fi/urn:nbn:fi:lb-2020031201>
- Sun, Z., Deng, Z., Nie, J., & Tang, J. (2019). Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. Retrieved from <https://openreview.net/forum?id=HkgEQnRqYQ>
- Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016). Complex embeddings for simple link prediction. In M. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016* (Vol. 48, pp. 2071–2080). JMLR.org. Retrieved from <http://proceedings.mlr.press/v48/trouillon16.html>
- Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., & Tang, J. (2021). Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9, 176-194. doi: 10.1162/tacl_a_00360
- Wu, F., Qiao, Y., Chen, J.-H., Wu, C., Qi, T., Lian, J., ... Zhou, M. (2020, July). MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3597–3606). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.331
- Yang, B., Yih, W., He, X., Gao, J., & Deng, L. (2015). Embedding entities and relations for learning and inference in knowledge bases. In Y. Bengio & Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Retrieved from <http://arxiv.org/abs/1412.6575>
- Zhang, C., Gao, W., Song, J., & Jiang, J. (2016). An imbalanced data classification algorithm of improved autoencoder neural network. In *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)* (p. 95-99). doi: 10.1109/ICACI.2016.7449810
- Zhang, S., Tay, Y., Yao, L., & Liu, Q. (2019). Quaternion knowledge graph embeddings. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada* (pp. 2731–2741). Retrieved from <https://proceedings.neurips.cc/paper/2019/hash/d961e9f236177d65d21100592edb0769-Abstract.html>
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.
- Zhu, Z., Xu, S., Qu, M., & Tang, J. (2019). Graphvite: A high-performance cpu-gpu hybrid system for node embedding. In *The World Wide Web Conference* (pp. 2494–2504).



- Zosa, E., Granroth-Wilding, M., & Pivovarova, L. (2020, May). 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)* (pp. 32–37). Marseille, France: European Language Resources Association. Retrieved from <https://www.aclweb.org/anthology/2020.clssts-1.6>
- Zosa, E., Pivovarova, L., Boggia, M., & Ivanova, S. (2022). Multilingual topic labelling of news topics using ontological mapping. In *44th European Conference on Information Retrieval*. (to appear)
- Škrlić, B., Repar, A., & Pollak, S. (2019). Rakun: Rank-based keyword extraction via unsupervised learning and meta vertex aggregation. *Lecture Notes in Computer Science*, 311–323. Retrieved from http://dx.doi.org/10.1007/978-3-030-31372-2_26 doi: 10.1007/978-3-030-31372-2_26

Appendix A: EMBEDDIA Tools, Datasets and Challenges: Resources and Hackathon Contributions

EMBEDDIA Tools, Datasets and Challenges: Resources and Hackathon Contributions

Senja Pollak Jožef Stefan Institute <i>senja.pollak@ijs.si</i>	Marko Robnik Šikonja University of Ljubljana	Matthew Purver Queen Mary University of London Jožef Stefan Institute
Michele Boggia University of Helsinki	Ravi Shekhar Queen Mary University of London	Marko Pranjić Trikoder d.o.o.
Salla Salmela Suomen Tietotoimisto STT	Ivar Krustok Tarmo Paju Ekspress Meedia	Carl-Gustav Linden University of Bergen
Leo Leppänen Elaine Zosa University of Helsinki	Matej Ulčar University of Ljubljana	Linda Freienthal Silver Traat TEXTA OÜ
Luis Adrián Cabrera-Diego University of La Rochelle, L3i	Matej Martinc Nada Lavrač Blaž Škrlić Jožef Stefan Institute	Martin Žnidaršič Andraž Pelicon Boshko Koloski Jožef Stefan Institute
Vid Podpečan Janez Kranjc Jožef Stefan Institute	Shane Sheehan Usher Institute University of Edinburgh	Emanuela Boros University of La Rochelle, L3i
Jose G. Moreno University of Toulouse, IRIT	Antoine Doucet University of La Rochelle, L3i	Hannu Toivonen University of Helsinki <i>hannu.toivonen@helsinki.fi</i>

Abstract

This paper presents tools and data sources collected and released by the EMBEDDIA project, supported by the European Union's Horizon 2020 research and innovation program. The collected resources were offered to participants of a hackathon organized as part of the EACL Hackathon on News Media Content Analysis and Automated Report Generation in February 2021. The hackathon had six participating teams who addressed different challenges, either from the list of proposed challenges or their own news-industry-related tasks. This paper goes beyond the scope of the hackathon, as it brings together in a coherent and compact form most of the resources developed, collected and released by the EMBEDDIA project. Moreover, it constitutes a handy source for news media industry and researchers in the fields of Natural Language Processing and Social Science.

1 Introduction

News media industry is the primary provider of information for society and individuals. Since the first newspaper was published, the propagation of information has continuously changed as new technologies are adopted by the news media, and the advent of the internet has made this change faster than ever (Pentina and Tarafdar, 2014). Internet-based media (e.g., social media, forums and blogs) have made news more accessible, and dissemination more affordable, resulting in drastically increased media coverage. Social media can also help provide source information for newsrooms, as shown in e.g., disaster response tasks (Alam et al., 2018).

Suitable Natural Language Processing techniques are needed to analyze news archives and gain insight about the evolution of our society, while dealing with the constant flow of information. Relevant datasets are equally important in

order to train data-driven approaches. To encourage the development and uptake of such techniques and datasets, and take on the challenges presented by the introduction of new technologies in the news media industry, the EMBEDDIA project¹ organized, in conjunction with EACL 2021, a hackathon² as part of the EACL Hackashop on News Media Content Analysis and Automated Report Generation³.

For this event, held virtually in February 2021, the datasets and tools curated and implemented by the EMBEDDIA project were publicly released and made available to the participants. We also provided examples of realistic challenges faced by today's newsrooms, and offered technical support and consultancy sessions with a news media expert throughout the entire duration of the hackathon.

The contributions of this paper are structured as follows. Section 2 presents the tools released for the event. The newly gathered, publicly released EMBEDDIA datasets are reported in Section 3. Section 4 presents sample news media challenges. Section 5 outlines the projects undertaken by the teams who completed the hackathon. The hackathon outcomes are summarized in Section 6.

2 Tools

The EMBEDDIA tools and models released for the hackathon include general text processing tools like language processing frameworks and text representation models (Section 2.1), news article analysis (Section 2.2), news comment analysis (Section 2.3), and news article and headline generation (Section 2.4) tools.

These tools require different levels of technical proficiency. Language processing tools and frameworks require little to no programming skills. On the other hand, for some tasks, we provide fully functional systems that can be used out of the box but require a certain level of technical knowledge in order to be fully utilized. Moreover, some tools and text representation models require programming skills and can be employed to improve existing systems, implement new analytic tools, or to be adapted to new uses.

¹<http://embeddia.eu>

²<http://embeddia.eu/hackashop2021-call-for-hackathon-participation/>

³<http://embeddia.eu/hackashop2021/>

2.1 General Text Analytics

We first present two general frameworks, requiring no programming skills: the EMBEDDIA Media Assistant, incorporating the TEXTA Toolkit that is focused exclusively on text, and the ClowdFlows toolbox, which is a general data science framework incorporating numerous NLP components. Finally, we describe BERT embeddings, a general text representation framework that includes variants of multilingual BERT models, which are typically part of programming solutions.

2.1.1 TEXTA Toolkit and EMBEDDIA Media Assistant

The TEXTA Toolkit (TTK) is an open-source software for building RESTful text analytics applications.⁴ TTK can be used for:

- searching and aggregating data (using e.g. regular expressions),
- training embeddings,
- building machine learning classifiers,
- building topic-related lexicons using embeddings,
- clustering and visualizing data, and
- extracting and creating training data.

The TEXTA Toolkit is the principal ingredient of the EMBEDDIA Media Assistant (EMA), which includes the TEXTA Toolkit GUI and API, an API Wrapper with a number of APIs for news analysis, and a Demonstrator for demonstrating the APIs.

2.1.2 ClowdFlows

ClowdFlows⁵ is an open-source online platform for developing and sharing data mining and machine learning workflows (Kranjc et al., 2012). It works online in modern Web browsers, without client-side installation. The user interface allows combining software components (called widgets) into functional workflows, which can be executed, stored, and shared in the cloud. The main aim of ClowdFlows is to foster sharing of workflow solutions in order to simplify the replication and adaptation of shared work. It is suitable for prototyping, demonstrating new approaches, and exposing solutions to potential users who are not proficient in programming but would like to experiment with their own datasets and different tool parameter settings.

⁴<https://docs.texta.ee/>

⁵<https://cf3.ijs.si/>

2.1.3 BERT Embeddings

CroSloEngual⁶ BERT and FinEst⁷ BERT (Ulčar and Robnik-Šikonja, 2020) are trilingual models, based on the BERT architecture (Devlin et al., 2019), created in the EMBEDDIA project to facilitate easy cross-lingual transfer. Both models are trained on three languages: one of them being English as a resource-rich language, CroSloEngual BERT was trained on Croatian, Slovenian, and English data, while FinEst BERT was trained on Finnish, Estonian, and English data.

The advantage of multi-lingual models over monolingual models is that they can be used for cross-lingual knowledge transfer, e.g., a model for a task for which very little data is available in a target language such as Croatian or Estonian can be trained on English (with more data available) and transferred to a less-resourced language. While massive multilingual BERT-like models are available that cover more than 100 languages (Devlin et al., 2019), a model trained on only a few languages performs significantly better on these (Ulčar and Robnik-Šikonja, 2020). The two trilingual BERT models here are effective for the languages they cover and for the cross-lingual transfer of models between these languages. The models represent words/tokens with contextually dependent vectors (word embeddings). These can be used for training many NLP tasks, e.g., fine-tuning the model for any text classification task.

2.2 News Article Analysis Tools

The majority of provided tools cover different aspects of news article analysis, processing, and generation. We present keyword extraction tools TNT-KID and RaKUn, named entity recognition approaches, tools for diachronic analysis of words, tools for topic analysis and visualization, and tools for sentiment analysis.

2.2.1 Keyword Extraction

Two tools are available for keyword extraction: TNT-KID and RaKUn.

TNT-KID⁸ (Transformer-based Neural Tagger for Keyword Identification, Martinc et al., 2020) is a supervised tool for extracting keywords from

news articles in several languages (English, Estonian, Croatian, and Russian). It relies on the modified Transformer architecture (Vaswani et al., 2017) and leverages language model pretraining on a domain-specific corpus. This gives competitive and robust performance while requiring only a fraction of the manually labeled data needed by the best performing supervised systems. This makes TNT-KID especially appropriate for less-resourced languages where large manually labeled datasets are scarce.

RaKUn⁹ (Škrlj et al., 2019) offers unsupervised detection and exploration of keyphrases. It transforms a document collection into a network, which is pruned to keep only the most relevant nodes. The nodes are ranked, prioritizing nodes corresponding to individual keywords and paths (keyphrases comprised of multiple words). Being unsupervised, RaKUn is well suited for less-resourced languages where expensive pre-training is not possible.

2.2.2 Named Entity Recognition¹⁰

The Named Entity Recognition (NER) system is based on the architecture proposed by Boros et al. (2020). It consists of fine-tuned BERT with two additional Transformer blocks (Vaswani et al., 2017). We provided models capable of predicting three types of named entities (Location, Organisation and Person) for eight European languages: Croatian, Estonian, Finnish, Latvian, Lithuanian, Russian, Slovene and Swedish. These models were trained using the WikiANN corpus (Pan et al., 2017), specifically using the training, development and testing partitions provided by Rahimi et al. (2019). Regarding BERT, for Croatian and Slovene we used *CroSloEngual BERT* (Ulčar and Robnik-Šikonja, 2020); for Finnish and Estonian *FinEst BERT* (Ulčar and Robnik-Šikonja, 2020); for Russian *RuBERT* (Kuratov and Arkhipov, 2019); for Swedish *Swedish BERT* (Malmsten et al., 2020); for Latvian and Lithuanian *Multilingual BERT* (Devlin et al., 2019).

2.2.3 Diachronic News Analysis¹¹

The tool for diachronic semantic shift detection (Martinc et al., 2019a) leverages the BERT contextual embeddings (Devlin et al., 2019) for generat-

⁶<https://huggingface.co/EMBEDDIA/crosloengual-bert>

⁷<https://huggingface.co/EMBEDDIA/finest-bert>

⁸https://github.com/EMBEDDIA/tnt_kid

⁹<https://github.com/EMBEDDIA/RaKUn>

¹⁰<https://github.com/EMBEDDIA/stacked-ner>

¹¹https://github.com/EMBEDDIA/semantic_shift_detection

ing time-specific word representations. It checks whether a specific word (or phrase) in the corpus has changed across time by measuring the rate of change for time-specific relations to semantically similar words in distinct time periods. Besides measuring long-term semantic changes, the method can also be successfully used for the detection of short-term yearly semantic shifts and has even been employed in the multilingual setting.

2.2.4 Topic Analysis

We present three tools dealing with news topics: PTM, PDTM and TeMoCo. The first two use topics to link articles across languages, and the third one visualizes distributions of topics over time.

PTM¹² (Polylingual Topic Model, [Mimno et al., 2009](#)) can be used to train cross-lingual topic models and obtain cross-lingual topic vectors for news articles. These vectors can be used to link news articles across languages. An ensemble of cross-lingual topic vectors and document embeddings can outperform stand-alone methods for cross-lingual news linking ([Zosa et al., 2020](#)).¹³

PDTM¹⁴ (Polylingual Dynamic Topic Model, [Zosa and Granroth-Wilding, 2019](#)) is an extension of the Dynamic Topic Model ([Blei and Lafferty, 2006](#)) for multiple languages. This model can track the evolution of topics over time aligned across multiple languages.

TeMoCo¹⁵ (Temporal Topic Visualisation, [Sheehan et al., 2019, 2020](#)) visualizes changes in topic distribution and associated keywords in a document or collection of articles. The tool can investigate a single document or a corpus which has been temporally annotated (e.g., a transcript or corpus of dated articles). The user can examine an overview of a dataset, processed into time and topic segments. The changes in topic size and keywords describe patterns in the data. Clicking on the segments brings up the related news articles with keyword highlighting.

¹²<https://github.com/EMBEDIA/cross-lingual-linking>

¹³<https://github.com/EMBEDIA/cross-lingual-linking>

¹⁴https://github.com/EMBEDIA/multilingual_dtm

¹⁵<https://github.com/EMBEDIA/TeMoCo>

2.2.5 News Sentiment Analysis¹⁶

Sentiment analysis is likely the most popular NLP application in industry. Our multilingual model for news sentiment classification is based on multilingual BERT. The model was trained on the Slovenian news sentiment dataset ([Bučar et al., 2018](#)) using a two-step training approach with document and paragraph level sentiment labels ([Pelicon et al., 2020](#)). The model was tested on the document-level labels of the Croatian news sentiment dataset (Section 3.2.2) in a zero-shot setting. The model maps the input document into one of the three predefined classes: positive, negative, and neutral.

2.3 News Comment Analysis Tools

Several of the tools in the sections above can also be applied to comments. We describe the following comment-specific tools: comment moderation, bot and gender detection, and sentiment analysis tools.

2.3.1 Comment Moderation¹⁷

Our comment moderation tool flags inappropriate comments that should be blocked from appearing on news sites ([Pelicon et al., 2021a,b](#)). It uses multilingual BERT ([Devlin et al., 2019](#)) and the trilingual EMBEDIA BERT models (Section 2.1.3). The models were trained on combinations of five datasets: Croatian and Estonian (see Section 3.3 and details in [Shekhar et al. \(2020\)](#)), Slovenian ([Ljubešić et al., 2019](#)), English ([Zampieri et al., 2019](#)), and German ([Wiegand et al., 2018](#)). For Croatian, we also provide a model to predict which rule is violated, based on the moderation policy of 24 sata, the biggest Croatian news publisher (see Section 3.3.3).

2.3.2 Bot and Gender Detection¹⁸

An author profiling tool for gender classification and bot detection in Spanish and English, trained on Twitter data ([Martinc et al., 2019b](#)), was developed for the PAN 2019 author profiling shared task ([Rangel and Rosso, 2019](#)). It uses a two-step approach: in the first step distinguishing between bots and humans, and in the second step determining the gender of human authors. It relies on a Logistic Regression classifier and employs a number of different word and character n-gram features.

¹⁶https://github.com/EMBEDIA/crosslingual_news_sentiment

¹⁷https://github.com/EMBEDIA/hackashop2021_comment_filtering

¹⁸<https://github.com/EMBEDIA/PAN2019>

2.3.3 Sentiment Analysis¹⁹

The code for sentiment analysis allows training a model that classifies text into one of three sentiment categories: positive, neutral, or negative. The classifier is trained on the Twitter datasets²⁰ provided by Mozetič et al. (2016). The models and datasets support cross-lingual knowledge transfer from resource-rich language(s) to less-resourced languages.

2.4 News Article and Headline Generation

Two of our tools are for generating text, either news for specific topics, or creative language.

Template-Based NLG System for Automated Journalism The rule-based natural language generation system—similar in concept to Leppänen et al. (2017)—produces news texts in Finnish and English from statistical data obtained from EuroStat. The system provides the text inputs used in the NLG challenges, described in Section 4.3. Access to the tool is provided through an API.²¹

Creative Language Generation We provide a framework²² to help in generation of creative language using an evolutionary algorithm (Alnajjar and Toivonen, 2020).

3 Datasets

For the purposes of the hackashop, the EMBEDDIA media partners released their news archives, the majority of which are now being made publicly available for use after the project.

3.1 General EMBEDDIA News Datasets

Four publicly available datasets released by the EMBEDDIA project are described below.

3.1.1 Ekspress Meedia News Archive (in Estonian and Russian)

Ekspress Meedia belongs to the Ekspress Meedia Group, one of the largest media groups in the Baltics. The dataset is an archive of articles from the Ekspress Meedia news site from 2009–2019, containing over 1.4M articles, mostly in the Estonian (1,115,120 articles) with some in the Russian

language (325,952 articles). Keywords (tags) are included for articles after 2015. The dataset is publicly available in the CLARIN repository.²³

3.1.2 Latvian Delfi Article Archive (in Latvian and Russian)

Latvian Delfi belongs to Ekspress Meedia Group. This dataset is an archive of articles from the Delfi news site from 2015–2019, containing over 180,000 articles (c. 50% in Latvian and 50% in Russian language). Keywords (tags) for articles are included. The dataset is publicly available in CLARIN.²⁴

3.1.3 24sata News Archive (in Croatian)

24sata is the biggest Croatian news publisher, owned by the Styria Media Group. The 24sata news portal consists of a daily news portal and several smaller portals covering news on specific topics, such as automotive news, health, culinary content, and lifestyle advice. The dataset contains over 650,000 articles in Croatian between 2007–2019, as well as assigned tags. The dataset is publicly available in CLARIN.²⁵

3.1.4 STT News Archive (in Finnish)

The Finnish corpus (STT, 2019) contains newswire articles in Finnish sent to media outlets by the Finnish News Agency (STT) between 1992–2018. The corpus includes about 2.8 million items in total. The news articles are categorized by department (domestic, foreign, economy, politics, culture, entertainment and sports), as well as by metadata (IPTC subject categories or keywords and location data). The dataset is publicly available via CLARIN,²⁶ as is a parsed version of the corpus in CoNLL-U format (STT et al., 2020).²⁷

3.2 Task-specific News Datasets

For the purposes of the hackashop, a set of task-specific datasets were also gathered.

3.2.1 Keyword Extraction Datasets

For the keyword extraction challenge, we created train and test data splits, given as article IDs from datasets in Section 3.1. The number of articles for Estonian, Latvian, Russian and Croatian (see Koloski et al. (2021a) for details) are:

²³<http://hdl.handle.net/11356/1408>

²⁴<http://hdl.handle.net/11356/1409>

²⁵<http://hdl.handle.net/11356/1410>

²⁶<http://urn.fi/urn:nbn:fi:lb-2019041501>

²⁷<http://urn.fi/urn:nbn:fi:lb-2020031201>

¹⁹<https://github.com/EMBEDDIA/cross-lingual-classification-of-tweet-sentiment>

²⁰<http://hdl.handle.net/11356/1054>

²¹<http://newseye-wp5.cs.helsinki.fi:4220/documentation/>

²²<https://github.com/EMBEDDIA/evolutionary-algorithm-for-NLG>

- Croatian: 32,223 train, 3,582 test;
- Estonian: 10,750 train, 7,747 test;
- Russian: 13,831 train, 11,475 test;
- Latvian: 13,133 train, 11,641 test.

The data is publicly available in CLARIN.²⁸

3.2.2 News Sentiment Annotated Dataset

We selected a subset of 2,025 news articles from the Croatian 24sata dataset (see Section 3.1.3 and Peli-con et al., 2020). Several annotators annotated the articles on a five-point Likert-scale from 1 (most negative sentiment) to 5 (most positive). The final sentiment label of an article was then based on the average of the scores given by the different annotators: negative if average was less than or equal to 2.4, neutral if between 2.4 and 3.6, or positive if greater than or equal to 3.6. The dataset is publicly available in CLARIN.²⁹

3.2.3 Estonian-Latvian Interesting News Pairs

For the purposes of the challenge on finding interesting news from neighbouring countries (see Section 4.1.2 and Koloski et al., 2021b) an Estonian journalist gathered 21 news articles from Latvia that would be of interest for Estonians, paired with 21 corresponding Estonian articles.³⁰

3.2.4 Corpus of Computer-Generated Statistical News Texts

This corpus, consisting of a total 188 news texts produced by the rule-based natural language generation system described in Section 2.4, is provided to allow for easier offline development of solutions to the NLG challenges. The corpus contains news texts in both Finnish and English,³¹ discussing consumer prices as well as health care spending and funding on the national level within the EU.

3.3 News Comments Datasets

Three news comment datasets have been made publicly available. To ensure privacy, user IDs in all news comment datasets in this section have been obfuscated, so they no longer correspond to the original IDs on the publishers' systems. User IDs for moderated comments have been removed.

²⁸<http://hdl.handle.net/11356/1403>

²⁹<http://hdl.handle.net/11356/1342>

³⁰<https://github.com/EMBEDDIA/interesting-cross-border-news-discovery>

³¹<https://github.com/EMBEDDIA/embeddia-nlg-output-corpus>

3.3.1 Ekspress Meedia Comment Archive (in Estonian and Russian)

This dataset is an archive of reader comments on the Ekspress Meedia news site from 2009–2019, containing approximately 31M comments, mostly in Estonian language, with some in Russian. The dataset is publicly available in CLARIN.³²

3.3.2 Latvian Delfi Comment Archive (in Latvian and Russian)

The dataset of Latvian Delfi, which belongs to Ekspress Meedia Group, is an archive of reader comments from the Delfi news site from 2014–2019, containing approximately 12M comments, mostly in Latvian language, with some in Russian. The dataset is publicly available in CLARIN.³³

3.3.3 24sata Comment Archive (in Croatian)

In this archive, there are over 20M user comments from 2007–2019, written mostly in Croatian. All comments were gathered from 24sata, the biggest Croatian news publisher, owned by Styria Media Group. Each comment is given with the ID of the news article where it was posted and with multi-label moderation information corresponding to the rules of 24sata's moderation policy (see Shekhar et al., 2020). The dataset is publicly available in CLARIN.³⁴

3.4 Other News Datasets

EventRegistry (Leban et al., 2014), which is a news intelligence platform aiming to empower organizations to keep track of world events and analyze their impact, provided free access to their data for hackathon participants.

Datasets relevant to the hackathon have also been made available for academic use by the Finnish broadcasting company Yle in Finnish³⁵ and in Swedish³⁶.

4 Challenges

Sample news media challenge addressed in the EMBEDDIA project come from three different areas: news analysis, news comments analysis, and article and headline generation.

³²<http://hdl.handle.net/11356/1401>

³³<http://hdl.handle.net/11356/1407>

³⁴<http://hdl.handle.net/11356/1399>

³⁵<https://korp.csc.fi/download/YLE/fi/2011-2018-src/>

³⁶<https://korp.csc.fi/download/YLE/sv/2012-2018-src/>

4.1 News Analysis Challenges

4.1.1 Keyword Extraction

The EMBEDDIA datasets from Ekspress Meedia, Latvian Delfi and 24sata contain articles together with keywords assigned by journalists (see Section 3.2.1). The project has produced several state-of-the-art approaches for automatic keyword extraction on these datasets (see Section 2.2.1). The challenge consists of providing alternative methods to achieve the most accurate keyword extraction and compare with our results.

4.1.2 Identifying Interesting News from Neighbouring Countries

Journalists are very interested in identifying stories from cross-border countries, that attract a large number of readers and are “special”. A journalist at Ekspress Meedia in Estonia gave the example of selecting news from Latvia that would be of interest to Estonian readers. Example topics include: drunk Estonians in Latvia, a person in Latvia living in a boat, stories from Latvia about topics that also interest Estonians (for example, alcohol taxes, newsworthy actions that take place near the border, certain public figures). At the moment it is easy to detect all the news from Latvia with the mentions of words “Estonia” or “Estonians”, but the challenge is to identify a larger number of topics, e.g. scandals, deaths, gossip that might be somehow connected to Estonia, and news and stories that Estonians relate to (for example, when similar things have happened in Estonia or similar news has been popular there). Given the collection of news from two different countries (e.g. Estonia, Latvia, see Section 3.1), the task is to identify these special interesting news stories; 21 manually identified examples were provided (see Section 3.2.3).

4.1.3 Diachronic News Article Analysis

Media houses with large news articles collections are interested in analysing the reporting on certain topics to investigate changes over time. This can not only help them understand their reporting, but also help journalists to discover specific aspects related to these concepts.

An example from a news media professional from Estonia is as follows: “the doping affairs in sports regularly appear and for example for one of our skiers, a few years ago, we have already reported on a potential doping affair, but did not analyse it in depth. Few years later it has turned out that the sportsman was indeed involved in a doping

affair. Having a better overview of doping related persons and topics over time, would be interesting for us.” An even more straightforward application is the monitoring of politicians and parties; controversial topics are also of interest, as they can show general changes in society towards them.

Each of the media partners provided some people/parties/concepts of their interest. Examples are reported in Appendix A.

4.2 News Comments Analysis

4.2.1 Comment Moderation

The EMBEDDIA datasets from Ekspress Meedia and 24sata contain comments with metadata showing the ones blocked by the moderators (see Section 3.3). In the case of the 24sata dataset, specific moderation policies exist with a list of reasons for blocking, and the metadata also shows which of the reasons applied. The policies are applied by humans, though, and therefore the metadata will reflect the way moderators actually behave, including making mistakes and showing biases. During the EMBEDDIA project, we have developed and evaluated multiple automatic filtering approaches on these datasets, which can be used off-the-shelf or can be re-trained or modified (see Section 2.3.1). The hackathon participants were invited to propose alternative comment filtering methods, to improve over the existing approaches, or apply them to other datasets; to use them to investigate how human moderators actually behave; and/or to investigate how to analyse, understand or use the outputs.

4.2.2 Comment Summarization

Each of the comment datasets available contains about 10 years of data. The EMBEDDIA project has developed and evaluated a range of classifiers that can detect useful information in comments and comment-like text (including sentiment, topic, author information etc; see Section 2.3). The participants were invited to use these and other methods to extract meaningful information from comment threads and develop new ways of presenting this information in a way that could be useful to a journalist or analyst. Example approaches given were summarizing topics, views and opinions; and detecting and summarizing constructive or positive comments, as an antidote to the negative comments so often focused on in NLP.

4.3 Natural Language Generation

4.3.1 Improving the Fluency of Automatically Generated Articles

Despite recent strides in neural natural language generation (NLG) methods, neural NLG methods are still prone to producing text that is not grounded in the input data. As such errors are catastrophic in news industry applications, most news generation systems continue to employ rule-based NLG methods. Such methods, however, lack to adequately handle the variety and fluency of expression. One potential solution would be to combine neural post-processing with a rule-based NLG system. In this challenge, participants are provided with black box access to a rule-based NLG system that produces statistical news articles. A corpus of the produced news articles is also provided.³⁷ The challenge is to use automated post-processing methods to improve the fluency and grammaticality of the system's output without changing the meaning of the text.

The system is multilingual (English and Finnish), and optimally the proposed solutions should be language-independent, taking advantage of e.g., multilingual word embeddings. At the same time, we also welcome monolingual solutions.

4.3.2 Headline Generation

Headlines play an important role in news text, not only summarizing the most important information in the underlying news text, but also presenting it in a light that is likely to entice the reader to engage with the larger text. In this challenge, the participants are invited to create headlines for automatically generated articles (see Section 4.3.1).

5 Hackathon Contributions

Six teams with 24 members in total participated in the hackathon during 1–19 February 2021. The challenges described in Section 4 were offered to the teams as examples of interesting problems in the area of news media analysis and generation. The teams had, however, the freedom to choose and formulate their own aims for the hackathon. Likewise, they were offered the data, tools and models described above.

The hackathon was organized online, with three joint events to kick off the activities, to meet and talk about the ongoing work halfway, and to wrap up the work at the end. Ample support on tools,

models, data and challenges was provided by the EMBEDDIA experts via several channels.

The six teams all picked up different challenges and set themselves specific goals. Reports from five teams are included in these proceedings.

Three teams worked on news content analysis:

- One team developed a COVID-19 news dashboard to visualise sentiment in pandemic-related news. The dashboard uses a multilingual BERT model to analyze news headlines in different languages across Europe (Robertson et al., 2021).
- Methods for cross-border news discovery were developed by another team using multilingual topic models. Their tool discovers Latvian news that could interest Estonian readers (Koloski et al., 2021b).
- A third team used sentiment and viewpoint analysis to study attitudes related to LGBTIQ+ in Slovenian news. Their results suggest that political affiliation of media outlets can affect sentiment towards and framing of LGBTIQ+-specific topics (Martinc et al., 2021).

Two teams looked at different challenges related to comment analysis:

- One team automated news comment moderation. They compiled and labeled a dataset of English news and social posts, and experimented with cross-lingual transfer of comment labels from English and subsequent supervised machine learning on Croatian and Estonian news comments (Korenčić et al., 2021).
- Another team looked at the diversity of news comment recommendations, motivated by democratic debate. They implemented a novel metric based on theories of democracy and used it to compare recommendation strategies of New York Times comments in English (Reuver and Mattis, 2021).

Finally, one team worked on a generation task:

- The team experimented with several methods for generating headlines, given the contents of a news story. They found that headlines formulated as questions about the story's content tend to be both informative and enticing.

³⁷<https://github.com/EMBEDDIA/embeddia-nlg-output-corpus>

6 Conclusions

This paper presents the contributions of the EMBEDDIA project, including a large variety of tools, new datasets of news articles and comments from the media partners, as well as challenges that were proposed to the participants of the EACL 2021 Hackathon on News Media Content Analysis and Automated Report Generation. The hackathon had six participating teams who addressed different challenges, either from the list of proposed challenges or their own news-industry-related tasks. In the future, the tools and resources described can be used for a large variety of new experiments, and we hope that the proposed challenges will be addressed by the wider NLP research community.

Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation program under grant 825153 (EMBEDDIA).

We would like to thank EventRegistry for providing free access to their data for hackathon participants.

References

- Firoj Alam, Ferda Ofli, Muhammad Imran, and Michael Aupetit. 2018. A Twitter tale of three hurricanes: Harvey, Irma, and Maria. *Proc. of ISCRAM, Rochester, USA*.
- Khalid Alnajjar and Hannu Toivonen. 2020. [Computational generation of slogans](#). *Natural Language Engineering*, First View:1–33.
- David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.
- Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, and Antoine Doucet. 2020. [Alleviating digitization errors in named entity recognition for historical documents](#). In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 431–441, Online. Association for Computational Linguistics.
- Joze Bučar, Martin Žnidarsic, and Janez Povh. 2018. Annotated news corpora and a lexicon for sentiment analysis in Slovene. *Language Resources and Evaluation*, 52:895–919.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Boshko Koloski, Senja Pollak, Blaž Škrlić, and Matej Martinc. 2021a. Extending neural keyword extraction with TF-IDF tagset matching. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.
- Boshko Koloski, Elaine Zosa, Timen Stepišnik-Perdih, Blaž Škrlić, Tarmo Paju, and Senja Pollak. 2021b. Interesting cross-border news discovery using cross-lingual article linking and document similarity. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.
- Damir Korenčić, Ipek Baris, Eugenia Fernandez, Katarina Leuschel, and Eva Sánchez Salido. 2021. To block or not to block: Experiments with machine learning for news comment moderation. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.
- Janez Kranjc, Vid Podpečan, and Nada Lavrač. 2012. ClowdFlows: A cloud based scientific workflow platform. In Peter A. Flach, Tijl Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 7524 of *Lecture Notes in Computer Science*, pages 816–819. Springer Berlin Heidelberg.
- Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *arXiv cs.CL*. Preprint: 1905.07213.
- Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 107–110.
- Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, and Hannu Toivonen. 2017. Data-driven news generation for automated journalism. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 188–197.
- Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English. In *International Conference on Text, Speech, and Dialogue*, pages 103–114. Springer.
- Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT. *arXiv cs.CL*. Preprint: 2007.01658.

- Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2019a. Leveraging contextual embeddings for detecting diachronic semantic shift. *arXiv preprint arXiv:1912.01072*.
- Matej Martinc, Nina Perger, Andraž Pelicon, Matej Ulčar, Andreja Vezovnik, and Senja Pollak. 2021. EMBEDDIA hackathon report: Automatic sentiment and viewpoint analysis of Slovenian news corpus on the topic of LGBTIQ+. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.
- Matej Martinc, Blaž Škrlić, and Senja Pollak. 2019b. Fake or not: Distinguishing between bots, males and females. In *CLEF (Working Notes)*.
- Matej Martinc, Blaž Škrlić, and Senja Pollak. 2020. Tnt-kid: Transformer-based neural tagger for keyword identification. *arXiv preprint arXiv:2003.09166*.
- David Mimno, Hanna Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 conference on Empirical Methods in Natural Language Processing*, pages 880–889.
- Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PLOS ONE*, 11(5):1–26.
- Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual Name Tagging and Linking for 282 Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.
- Andraž Pelicon, Marko Pranjić, Dragana Miljković, Blaž Škrlić, and Senja Pollak. 2020. Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, 10(17):5993.
- Andraž Pelicon, Ravi Shekhar, Matej Martinc, Blaž Škrlić, Matthew Purver, and Senja Pollak. 2021a. Zero-shot cross-lingual content filtering: Offensive language and hate speech detection. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*.
- Andraž Pelicon, Ravi Shekhar, Blaž Škrlić, Matthew Purver, and Senja Pollak. 2021b. Investigating cross-lingual training for offensive language detection. *Submitted, to appear*.
- Iryna Pentina and M. Tarafdar. 2014. From "information" to "knowing": Exploring the role of social media in contemporary news consumption. *Comput. Hum. Behav.*, 35:211–223.
- Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. *Massively Multilingual Transfer for NER*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.
- Francisco Rangel and Paolo Rosso. 2019. Overview of the 7th author profiling task at pan 2019: bots and gender profiling in Twitter. In *Working Notes Papers of the CLEF 2019 Evaluation Labs Volume 2380 of CEUR Workshop*.
- Myrthe Reuver and Nicolas Mattis. 2021. Implementing evaluation metrics based on theories of democracy in news comment recommendation (Hackathon report). In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.
- Frankie Robertson, Jarkko Lagus, and Kaisla Kajava. 2021. A COVID-19 news coverage mood map of Europe. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.
- Shane Sheehan, Pierre Albert, Masood Masoodian, and Saturnino Luz. 2019. TeMoCo: A visualization tool for temporal analysis of multi-party dialogues in clinical settings. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 690–695. IEEE.
- Shane Sheehan, Saturnino Luz, Pierre Albert, and Masood Masoodian. 2020. TeMoCo-Doc: A visualization for supporting temporal and contextual analysis of dialogues and associated documents. In *Proceedings of the International Conference on Advanced Visual Interfaces, AVI '20*, New York, NY, USA. Association for Computing Machinery.
- Ravi Shekhar, Marko Pranjić, Senja Pollak, Andraž Pelicon, and Matthew Purver. 2020. Automating News Comment Moderation with Limited Resources: Benchmarking in Croatian and Estonian. *Journal for Language Technology and Computational Linguistics (JLCL)*, 34(1).
- Blaž Škrlić, Andraž Repar, and Senja Pollak. 2019. Rakun: Rank-based keyword extraction via unsupervised learning and meta vertex aggregation. In *Statistical Language and Speech Processing*, pages 311–323, Cham. Springer International Publishing.
- STT. 2019. Finnish news agency archive 1992-2018, source (<http://urn.fi/urn:nbn:fi:lb-2019041501>).
- STT, Helsingin yliopisto, and Khalid Alnajjar. 2020. Finnish News Agency Archive 1992-2018, CoNLL-U, source (<http://urn.fi/urn:nbn:fi:lb-2020031201>).
- Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT. In *International*

- Conference on Text, Speech, and Dialogue*, pages 104–111. Springer.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.
- Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of the GermEval 2018 Workshop (GermEval)*.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*, pages 1415–1420.
- Elaine Zosa and Mark Granroth-Wilding. 2019. **Multilingual dynamic topic model**. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1388–1396, Varna, Bulgaria. INCOMA Ltd.
- Elaine Zosa, Mark Granroth-Wilding, and Lidia Pivovarov. 2020. A comparison of unsupervised methods for ad hoc cross-lingual document retrieval. In *Proceedings of the LREC 2020 Workshop on Cross-Language Search and Summarization of Text and Speech*. European Language Resources Association (ELRA).
- Estonian: Jüri Ratas, Kersti Kaljulaid, Kaja Kallas, Martin Helme
 - Croatian: Andrej Plenković (the prime minister), Zoran Milanović (the president), Kolinda Grabar-Kitarović (previous president), Milan Bandić (mayor of Zagreb)
- Interesting topics** were selected for all three languages to allow also cross-lingual comparisons:
- **corona crisis, pandemics:** Estonian: Koroonakriis, pandeemia; Finnish: korona, koronakriisi, pandemia, koronapandemia; Croatian: korona, koronavirus, korona kriza, pandemija, korona pandemija
 - **same sex rights, registered partnership act, marriage referendum:** Estonian: samasooliste õigused, kooseluseadus, abielureferendum; Finnish: tasa-arvoinen avioliitto, rekisteröity parisuhde; Croatian: referendum o braku, životno partnerstvo, civilno partnerstvo
 - **financial knowledge, savings, investing, pension:** Estonian: rahatarkus, säästmise, investeerimine, pension; Finnish: sijoittaminen, piensijoittaja, säästäminen, eläke, eläkkeet; Croatian: ulaganje, investiranje, mali ulagači, dionice, ušteđevina, mirovina, penzija
 - **doping:** same word in Estonian/Finnish/Croatian.

A Entities of Interest for Diachronic News Article Analysis Challenge

For the challenge described in Section 4.1.3, each of the media partners provided some people/parties/concepts of their interest. These include the following.

Political parties:

- Estonian (Eskpress meedia): Reformierakond, EKRE, Keskerakond
- Finnish (STT)³⁸: Suomen Sosialidemokraattinen Puolue, demarit, SDP, (sd.); Kokoomus, (kok.); Keskusta, (kesk.); Perussuomalaiset, (ps.); Kristillisdemokraatit, KD, (kd.)
- Croatian: Hrvatska demokratska zajednica (HDZ), Socijaldemokratska partija Hrvatske (SDP), Hrvatska narodna stranka (HNS), Most nezavisnih lista (MOST)

Popular people:

³⁸The names without brackets are names the parties use and the abbreviation inside brackets is the way to mark a mp's / other person's political party within a text. For example Jussi Halla-aho (ps.) said that-

Appendix B: RuShiftEval: a shared task on semantic shift detection for Russian

RuShiftEval: a shared task on semantic shift detection for Russian

Pivovarova Lidia
University of Helsinki
Finland

lidia.pivovarova@helsinki.fi

Kutuzov Andrey
University of Oslo
Norway

andreku@ifi.uio.no

Abstract

We present the first shared task on diachronic word meaning change detection for the Russian. The participating systems were provided with three sub-corpora of the Russian National Corpus — corresponding to pre-Soviet, Soviet and post-Soviet periods respectively — and a set of approximately one hundred Russian nouns. The task was to rank those nouns according to the degrees of their meaning change between periods.

Although RuShiftEval is in many respects similar to the previous tasks organized for other languages, we introduced several novel decisions that allow for using novel methods. First, our manually annotated semantic change dataset is split in more than two time periods. Second, this is the first shared task on word meaning change which provided a training set.

The shared task received submissions from 14 teams. The results of RuShiftEval show that a training set could be utilized for word meaning shift detection: the four top-performing systems trained or fine-tuned their methods on the training set. Results also suggest that using linguistic knowledge could improve performance on this task. Finally, this is the first time that contextualized embedding architectures (XLM-R, BERT and ELMo) clearly outperform their static counterparts in the semantic change detection task.

Keywords: semantic change detection, Russian, shared task

DOI: 10.28995/2075-7182-2021-20-XX-XX

RuShiftEval: соревнование по детектированию семантических сдвигов в русском языке

Пивоварова Лидия
Университет Хельсинки
Финляндия
lidia.pivovarova@helsinki.fi

Кутузов Андрей
Университет Осло
Норвегия
andreku@ifi.uio.no

Аннотация

Мы представляем первую дорожку по автоматическому определению изменения значений слов для русского языка. Участники дорожки получили три подкорпуса НКРЯ - досоветский, советский и постсоветский - и список из около ста русских существительных. Задача состояла в ранжировании этих слов по степени семантического сдвига между этими периодами.

Наша дорожка во многих отношениях похожа на предыдущие подобные соревнования, которые организовывались для других языков. Однако мы предложили несколько нововведений, которые позволили участникам протестировать новые подходы к этой задаче. Во-первых, мы опубликовали новый датасет, в котором данные разбиты более чем на два периода. Во-вторых, это первая дорожка по автоматическому определению семантических сдвигов, где участникам был предоставлен тренировочный набор данных.

Дорожка получила более сотни решений от четырнадцати участников. Результаты соревнования продемонстрировали полезность тренировочных данных для определения семантических сдвигов: четыре лучших результата были продемонстрированы моделями, которые тренировались или донастраивались на тренировочных данных. Результаты так же демонстрируют, что использование априорных лингвистических знаний или сложных языковых моделей улучшают показатели в этой задаче.

Ключевые слова: диахронические семантические сдвиги, детектирование семантических изменений

1 Introduction

Words change their semantics over time as a result of combination of various processes that affect language simultaneously. Automatic detection and measuring the degree of meaning change could accelerate research in the history of language and also support a number of text analysis tasks such as information retrieval or media monitoring.

The RuShiftEval shared task is aimed at the comparison of various methods for detection of word meaning shift from diachronic corpora. Recently, two shared tasks for semantic change detection were organized: SemEval Task 1 for English, German, Swedish and Latin [17], and DIACR-Ita for Italian [2]. RuShiftEval is the first attempt to organize such an event with Russian data.

In many aspects, we follow the practices established during the previous shared tasks. However, we introduced several novelties: first, we deal with *three* time periods, namely pre-Soviet, Soviet and post-Soviet; second, we provided the participants with a *training dataset*, thus allowing for using supervised methods.

The shared task is collocated with Dialogue 2021, the 27th International Conference on Computational Linguistics and Intellectual Technologies. The test and development datasets used in RuShiftEval are now publicly available, as well as the evaluation code and the baseline.¹

2 Related work

Automatic detection of word meaning change is a fast developing research area. The majority of modern approaches utilize distributional *word embeddings* to detect changes in word context over time. Overview of various approaches for this task could be found in the recent surveys [18, 4, 22].

To perform numerical evaluation, the problem is most commonly formulated as following: an input is a *corpus* split into several (usually two) time periods and a *set of words*; the task is to *rank* these words according to the degree of meaning change they have undergone between the periods. The performance is measured by rank correlation between a produced ranking and the gold manually created ranking. Alternatively, the task could be cast as binary classification of words into changed and not-changed classes. In this case, evaluation is also done as comparison against manual annotation.

Thus, manually annotated datasets are key components for development of lexical semantic change models. Since word meaning shift is a *lexicon-level phenomenon*, annotation should take into account many word usages from each periods, making it a time-consuming task. The most recent DUREL framework solves this by annotating pairs of sentences and then computing an averaged metric that generalizes these annotations [16]. We follow this approach in our shared task.

The first shared task on word meaning change detection was organized in 2020 as a part of SemEval conference (SemEval 2020 Task 1). The shared task [17] provided datasets for four languages — English, German, Swedish, and Latin — with several dozens manually annotated words for each language. The task included two subtasks, described above: binary classification and ranking. More than twenty teams participated in it. One of the main results of SemEval 2020 Task 1 was that type-base (static) embeddings are more suitable for *unsupervised* semantic shift detection than more recent contextualized embeddings currently dominating almost all other NLP tasks. Another important observation is a high variety across corpora: a method that yields the best performance for one corpus may not be the best for another one. Another shared task was organized for Italian [2], where the task was binary classification, and the results largely replicated those from the SemEval.

Although RuShiftEval is the first shared task on word meaning change for Russian, semantic shift detection methods have been previously applied to this language, e.g. in [10, 20]. This research is accelerated by publishing of time-specific sub-corpora of the Russian National Corpus (RNC), consisting of sentences from the texts created in the pre-Soviet, Soviet and post-Soviet time periods. Together they cover nearly full RNC.² It is important to note that the RuShiftEval organizers are fully aware that 1)

¹https://github.com/akutuzov/rushifteval_public

²The sentence-shuffled version of the RNC split into 3 sub-corpora corresponding to the RuShiftEval time periods was made freely available specifically for this shared task (it is required to sign a license agreement to get access to the corpora): <https://rusvectors.org/static/corpora/>

the division of Russian language history into these particular periods is not the only possible option and the boundaries could be drawn differently; 2) the RNC itself is not fully representative of the history of Russian. However, some decisions had to be made with respect to the time bin boundaries; the division we chose is at least motivated with regards to historical events and yields sub-corpora of comparable sizes. In the same vein, no Russian corpus other than the RNC is available which is large enough, covers long enough time span, and provides the creation dates for the texts.

These diachronic sub-corpora of the RNC have previously been already used to create the *RuSemShift* dataset [14], which includes two subsets, each of 70 words, manually annotated and ranked according to their change from pre-Soviet to Soviet and from Soviet to post-Soviet times respectively. For the *RuShiftEval* data annotation, we used the same corpora and followed the same annotation procedure, so *RuSemShift* could be used as a training set by task participants. However, two parts of the *RuSemShift* dataset use different sets of words, while for the shared task we use the same list of words for *all three periods*, in principle allowing to study continuous word sense dynamic across time.

3 Task overview

The shared task focuses on three time periods, naturally stemming from the history of the Russian language and society. The boundary years of 1917 and 1991 were omitted from the annotation due to their transitioning nature:

1. pre-Soviet (1700-1916);
2. Soviet (1918-1990);
3. post-Soviet (1992-2016).

The *RuShiftEval* dataset consists of 111 Russian nouns (99 in the test set and 12 in the development set), manually annotated with the degrees of their meaning change in three time period pairs:

1. between pre-Soviet and Soviet periods (so called *RuSemShift1* score);
2. between Soviet and post-Soviet periods (so called *RuSemShift2* score);
3. between pre-Soviet and post-Soviet periods (so called *RuSemShift3* score).

We did not rely on any assumption on the dependencies of these three scores and annotated all pairs independently. Note that the resulting *RuShiftEval* dataset (about 30 000 human judgments in total) is described in more detail in a separate paper [9], so it is only briefly presented here. As per reviewers' suggestions, we provide the full list of target words with their change scores in the Appendix (although we strongly recommend to use the maintained version in our GitHub repository).

The annotation was conducted using crowd-sourcing (Yandex.Toloka platform). It followed the DuReL workflow described in [16]. An annotator had to read and score two sentences containing a target word and belonging to different time periods. The sentences were randomly sampled from the corresponding sub-corpora of the Russian National Corpus. The scores (from 1 to 4) grade semantic relatedness between the target word meanings in two sentences. The 1 score denotes 'the senses are unrelated', and the 4 score denotes 'the senses are identical'.

Then individual scores were accumulated into mean semantic relatedness between word usages from two different time periods; this measure is also known as COMPARE and was introduced in [16]. Basically, it reflects human judgments about such relatedness averaged across about 30 sentence pairs containing the target word. Thus, the lower is the score (the closer it is to 1), the stronger is the degree of semantic change. For each sentence pair, the score was in turn averaged across at least 3 human annotators.

As has been mentioned in Section 2, the *RuSemShift* dataset [14] could be used for training (or simply for sanity check in the *Practice* phase), and we encouraged participants to do this. To find out whether using training data actually helps semantic change detection was one of the purposes of the *RuShiftEval* shared task. We can now confirm that the answer is positive; see Section 5 for details.

We recommended the participants to use the RNC for their data-driven solutions, since this corpus has been used to annotate the data. They were free to employ any other linguistic sources, and some actually did; again, see Section 5. Submissions of the participants were processed, evaluated and ranked with the

help of Codalab platform.³

To help participants to start with the task, we also provided static word embeddings pre-trained on diachronic sub-corpora of the RNC, using the CBOW algorithm [11], with context window size 5 and vector size 300. Each model was published in two variants: trained on raw tokens and trained on lemmas with part of speech tags ('ЗАВОД_NOUN', etc). These embeddings were used in the baseline solution, which was available as a part of the starting kit for the participants.

4 Evaluation workflow

The task was formulated as a ranking problem, similar to Subtask 2 of the SemEval 2020 Task 1 [17]: a set of Russian words should be ranked according to the strength of their meaning change. Thus, we did not make any binary decisions on whether a word has changed its meaning or not.

Importantly, it was one and the same set of words, for which the participants had to provide 3 semantic change scores per each word. The lower score meant a stronger change; the higher score meant a higher semantic similarity between word usages in different time periods, and thus a weaker change.

During the main *Evaluation* phase (February 22 - March 1, 2021), the participants were provided with a set of 99 target Russian words. For each word, they had to submit three non-negative values, corresponding to semantic change in the aforementioned time period pairs. These values were used to build 3 column-wise rankings: so called *RuSemShift1*, *RuSemShift2* and *RuSemShift3*. Since rank correlation was used as the evaluation metrics, the absolute numerical values of semantic change scores did not matter (only their relative ranks).

During the *Development* phase (February 1 - February 22, 2021), a small development set was provided (12 manually annotated Russian words), and the participants could submit their predictions to get a preliminary estimation of their system performance (no gold labels were openly published).

Before February 1, the shared task was in the *Practice* phase: the participants could submit predictions to the words from the *RuSemShift* test set [14]. This dataset was already public, so the true labels were known to everyone. This phase could be used to sanity check submission routines. There were only two time period pairs, each with its own set of words (this is how *RuSemShift* is built). We remind that in the *Development* and *Evaluation* phases, the participants had *one* set of words and *three* time period pairs.

Each participating team was able to submit up to 10 answers in the Evaluation phase, and up to 1000 answers in the Development phase. Submissions were evaluated using Spearman rank correlation between word ranking produced by a system and a gold ranking obtained in manual annotation. Thus, for each system we computed three correlations, for each of the time period pairs. The final ranking of the systems is based on averaging of the three scores.

5 Shared task results

In the Evaluation phase, we received submissions from 14 users (some of them in 4 different teams). Table 1 shows the performance of top submissions from each user or team (we give the name of the team by default or the name of the individual participant, if no team was associated with this submission). The teams are ranked by their average scores.

Some initial comments are due with regards to this table:

1. The baseline solution employed lemmatized diachronic embeddings trained on the Russian National Corpus⁴ and the simple local neighborhood method from [5].
2. The differences between the first and the second best performing systems are not statistically significant according to the Fisher test; the differences between the second and the third systems are statistically significant at $p = 0.06$ for *RuSemShift1* only. However, the differences between the top three systems and the rest of the submissions are all statistically significant.
3. Using median score instead of average score does not substantially change the ranking.

³<https://competitions.codalab.org/competitions/28340>

⁴These embeddings and diachronic corpora were available to all participants.

	Team	RuSemShift1	RuSemShift2	RuSemShift3	Mean	Type
1	GlossReader	0.781	0.803	0.822	0.802	token
2	DeepMistake	0.798	0.773	0.803	0.791	token
3	vanyatko	0.678	0.746	0.737	0.720	token
4	aryzhova	0.469	0.450	0.453	0.457	token
5	Discovery	0.455	0.410	0.494	0.453	token
6	UWB	0.362	0.354	0.533	0.417	type
7	dschlechtweg	0.419	0.373	0.383	0.392	type
8	jenskaiser	0.430	0.310	0.406	0.382	token
9	SBX-HY	0.388	0.281	0.439	0.369	type
	Baseline	0.314	0.302	0.381	0.332	type
10	svart	0.163	0.223	0.401	0.262	type
11	BykovDmitrii	0.274	0.202	0.307	0.261	token
12	fdzr	0.217	0.251	0.065	0.178	type

Table 1: Evaluation phase leaderboard (Spearman rank correlations). The Type column shows the type of the used distributional embeddings.

4. Bold names denote teams or individual participants who submitted papers with the description of their systems. For other participants, we rely on the contents of the ‘Description’ field in their Codalab submissions.
5. The DeepMistake team made several submissions of essentially the same system with varying hyperparameters; we show only the best one.
6. The SBX-HY team made a minor technical mistake, and their correlation scores were negative. Our opinion is that this does not undermine the developed system itself, so we show the absolute values in Table 1, and rank submissions accordingly.

5.1 Participating systems overview

Below, we give the descriptions of the participating systems. First, let us look at the submissions described in the submitted papers.

GlossReader [13] relied on the pretrained multilingual XLM-R language model [21]. On top of it, they trained a word sense disambiguation (WSD) system on English WSD datasets, using learned representations of sense definitions. Interestingly, this system shows excellent performance on Russian lexical semantic change data as well. Essentially, this participant reproduced the RuShiftEval annotation effort, replacing human judgments with the distances between XML-R contextualized embeddings of the target words. Additionally, a linear regression was trained on the *RuSemShift* dataset to convert vector distance values into relatedness scores (from 1 to 4).

DeepMistake [3] used the multilingual XLM-R as well, and also pre-trained on English WSD datasets, but without explicitly predicting senses. Similarly to **GlossReader**, they additionally fine-tuned this model on the *RuSemShift* using linear regression for mapping to relatedness scores.

aryzhova [15] tried both ruBERT [7] and ELMo contextualized embeddings.⁵ Interestingly, in their experiments ELMo outperformed BERT. Note, however, that **aryzhova** system is different from **vanyatko** (described below) in that it does not fine-tune BERT or ELMo: instead, it calculates the average cosine similarity between target word embeddings (sometimes with the addition of the neighboring word

⁵The ELMo models for Russian were borrowed from the RusVectōrēs service.

tokens) in the sampled sentence pairs, reproducing the *APD* method from [8]. Another interesting experiment reported in the paper from this participant is using ‘grammatical vectors’ corresponding to the frequencies of 12 morphological forms of Russian nouns (6 cases and singular/plural forms). They report that the cosine similarities between such vectors calculated on different time bins improved the performance of relatedness score classifier (trained and evaluated on the *RuSemShift* dataset).

UWB [12] this team employed traditional 300-dimensional static word embedding (in particular, fast-Text). Orthogonal Procrustes and Canonical Correlation Analysis (CCA) were used for alignment, with CCA showing somewhat better results. The semantic change score was calculated as simple cosine similarity between word vectors across different time periods.

SBX-HY [6] again used static word embeddings, but in this case instead of post-training alignment, they relied on Temporal Referencing approach [19], successfully used for semantic change detection with other languages. In this approach, the target words are augmented with time period labels, and then one embedding model is trained on all available data. Hyper-parameters were selected based on the *RuSemShift* dataset. Interestingly, with the *RuShiftEval* data, Temporal Referencing barely managed to outperform the organizers’ baseline, which is an interesting negative result.

BykovDmitrii [1] employed an interesting approach with lexical substitutes produced by the multilingual XLM-R as a masked language model. These substitutes were then clustered into senses and the divergence between clusters from different time periods was used as the semantic change score. This particular approach failed, but in the post-evaluation phase, the participant managed to significantly improve their result by skipping the clustering step and instead directly comparing bags of lexical substitutes (see more in their paper).

Now let us briefly describe the systems which did not submit papers, based on their descriptions in CodaLab. **Vanyatko** employed the RuBERT model. They fine-tuned RuBERT with sentence pairs as inputs and relatedness scores (from 1 to 4) as outputs. Similar to **GlossReader** and **DeepMistake**, **vanyatko** tried to reproduce human annotation process. The **Discovery** team used BERT with ensemble of Average Pairwise Cosine Distance and Cosine Distance of averaged embeddings. **Dschlechtweg** trained regression on the labeled training examples from *RuSemShift* with SGNS embeddings. **Jenskaiser** also employed static SGNS embeddings and Temporal referencing or ‘word injection (WI)’. They got results very similar to **SBX-HY**. Finally, **svart** used orthogonal Procrustes and cosine distances with the lemmatized word2vec embeddings provided by the organizers, and **fdzr** again relied on temporal referencing.

6 Discussion

We believe the results of the *RuShiftEval* are interesting for the lexical semantic change detection field in at least four aspects.

1 This is the first time the systems based on *contextualized embeddings* top the leaderboard. In both SemEval 2020 Task 1 [17] and DIACR-ITA [2], type embedding (or ‘static’ embedding) based architectures clearly won the rankings. But at the *RuShiftEval*, five top performing systems use pre-trained contextualized (‘token-based’) models: XLM-R, BERT and ELMo. In the previous work, the researchers in the field expressed doubts about the abilities of token embeddings with relation to semantic change detection. It seems that at least in the case of *RuShiftEval*, they are perfectly able to solve the task better than their static counterparts. However, the best performing teams introduced completely novel approaches to the problem, so the distinction between our results and results of the previous tasks lies in the difference between models rather than between embeddings themselves.

2 Surprisingly, the first and the second best submissions relied on the contextualized XLM-R model [21], which was not even specifically trained for processing Russian data. Its training corpus included texts in about 100 languages. Russian is well represented there but is far from being the largest in absolute size. The results of our shared task show that multilingual models like XLM-R can be very

successfully applied to semantic change detection for Russian (and arguably for many other languages): their transferability is extremely high.

Interestingly, at the SemEval 2020 Task 1, the attempts to use XLM-R did not end up very well: the system based on it ended up 7th in the Subtask 2 (closest to RuShiftEval), well below the type-based architectures. One of the reasons for this can be the next insightful outcome of RuShiftEval:

3 Using training data helps lexical semantic change detection. As already said, the *RuSemShift* dataset [14] was publicly available by the beginning of RuShiftEval, and the participants were free to use it as they saw fit. The annotation procedures were identical for *RuSemShift* and the shared task test sets. Thus, one of the aims of RuShiftEval was to find out whether using previously annotated data can improve the performance of semantic change ranking. As it turns out, it definitely can. Four top systems all train or fine-tune on *RuSemShift*. This was the first semantic change detection shared task to introduce such a setup. At the same time, using unsupervised methods with parameters fine tuning on the training set does not seem to be a productive strategy.

4 Finally, at least two participants (both in the top of the leaderboard) used explicit linguistic knowledge in addition to statistical distributional models. In particular, **GlossReader** (the winner of the task) fine-tuned their XLM-R model to select a definition (a gloss) from the WordNet, that is most appropriate for a particular target word occurrence [13]. Note that it was not the plain old classification: the model directly processed the definitions themselves as sequences of words. Another example is **aryzhova** who employed a linguistic intuition that semantic change is often linked to fluctuations in the frequency of different grammatical forms [15]. We believe using linguistic knowledge is an interesting direction for future development of the semantic change detection field.

It is important to note that the observations above are applicable only to the shared task setup used in RuShiftEval: that is, ranking words by the degree of semantic change estimated with the COMPARE measure calculated on human annotations conducted within the DUREL framework. Actually, many of the top-performing systems essentially reproduced the annotation process with large language models, which seems to be successful even though they could not know which particular sentences were sampled for manual annotation. With other evaluation setups, different approaches could be at the top. As an example, it is known that the COMPARE measure is much influenced by sense frequencies and can easily overlook changes occurring to rare senses — either their appearance or disappearance. If the systems were evaluated based on explicit senses they managed to detect, clustering-based approaches would arguably rank much higher.

7 Conclusion

In this paper, we summarized the outcome of RuShiftEval: the first shared task on lexical semantic change detection for Russian. The purpose of the shared task was twofold: first, to evaluate current state-of-the-art methods in semantic change detection on Russian data, and second, to explore the possibilities of *supervised* semantic change detection. This was ensured by the prior existence of *RuSemShift* dataset, annotated in exactly the same way as our testing data.

The results of the shared task show that training on existing semantic change data is indeed useful and can significantly boost evaluation scores. In absolute values, the correlations with human judgments achieved by the RuShiftEval participants are much higher than those demonstrated in the SemEval 2020 Task 1 across English, Latin, German and Swedish (the best system there yielded 0.527). Note that although *RuSemShift* (used as a training set) and RuShiftEval (used as a development and a test set) are annotated similarly, they are not splits of one and the same dataset. Thus, we believe this finding to be reliable and expect it to hold for other languages as well.

Another interesting outcome of RuShiftEval is the strong victory of contextualized (token-based) embedding architectures over static (type-based) ones. This is different from the results of previous shared tasks on semantic change detection, and we believe this means the community has finally learned how to properly use contextualized embeddings for this task. This is even more impressive considering the fact that the winning systems used the multilingual XLM-R instead of a Russian-specific model.

Despite these substantial findings, our shared task has just started to pave the way for studying approaches to automatic semantic change detection in Russian. Our evaluation setup (ranking by aggregated COMPARE score) cannot capture the entire spectrum of semantic change. This linguistic phenomenon is extremely complex, and we are hoping that future shared tasks will try to account for that.

Acknowledgments

The annotation effort for this shared task was supported by the Russian Science Foundation grant 20-18-00206. We are especially grateful to Valery Solovyev (Kazan Federal University). This work has been partially supported by the European Union Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDDIA).

References

- [1] Arefyev Nikolay, Bykov Dmitrii. An Interpretable Approach to Lexical Semantic Change Detection with Lexical Substitution // Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue. — 2021.
- [2] DIACR-Ita@ EVALITA2020: Overview of the EVALITA2020 diachronic lexical semantics (DIACR-Ita) task / Pierpaolo Basile, Annalina Caputo, Tommaso Caselli et al. // Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), Online. CEUR. org. — 2020.
- [3] DeepMistake: Which Senses are Hard to Distinguish for a Word-in-Context Model / Nikolay Arefyev, Maksim Fedoseev, Vitaly Protasov et al. // Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue. — 2021.
- [4] Diachronic word embeddings and semantic shifts: a survey / Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, Erik Velldal // Proceedings of the 27th International Conference on Computational Linguistics. — 2018. — P. 1384–1397.
- [5] Hamilton William L., Leskovec Jure, Jurafsky Dan. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. — Austin, Texas : Association for Computational Linguistics, 2016. — Nov. — P. 2116–2121. — Access mode: <https://www.aclweb.org/anthology/D16-1229>.
- [6] Hengchen Simon, Viorica Kate, Indukaev Andrey. SBX-HY at RuShiftEval 2021: // Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue. — 2021.
- [7] Kuratov Yury, Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language // Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue. — 2019. — Access mode: <http://www.dialog-21.ru/media/4606/kuratovplusarkhipovm-025.pdf>.
- [8] Kutuzov Andrey, Giulianelli Mario. UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection // Proceedings of the Fourteenth Workshop on Semantic Evaluation. — Barcelona (online) : International Committee for Computational Linguistics, 2020. — Dec. — P. 126–134. — Access mode: <https://www.aclweb.org/anthology/2020.semeval-1.14>.
- [9] Kutuzov Andrey, Pivovarova Lidia. Three-part diachronic semantic change dataset for Russian // Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change. — online : Association for Computational Linguistics, 2021.
- [10] Measuring Diachronic Evolution of Evaluative Adjectives with Word Embeddings: the Case for English, Norwegian, and Russian / Julia Rodina, Daria Bakshandaeva, Vadim Fomin et al. // Proceedings of the 1st International Workshop on Computational Approaches to Historical Language

- Change. — Florence, Italy : Association for Computational Linguistics, 2019. — Aug. — P. 202–209. — Access mode: <https://www.aclweb.org/anthology/W19-4725>.
- [11] Mikolov Tomas, Yih Wen-tau, Zweig Geoffrey. Linguistic Regularities in Continuous Space Word Representations // Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Atlanta, Georgia : Association for Computational Linguistics, 2013. — Jun. — P. 746–751. — Access mode: <https://www.aclweb.org/anthology/N13-1090>.
- [12] Priban Pavel, Pražák Ondřej, Taylor Stephen. UWB@RuShiftEval: Measuring Semantic Difference as per-word Variation in Aligned Semantic Spaces // Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue. — 2021.
- [13] Rachinskiy Maxim, Arefyev Nikolay. Zero-shot Cross-lingual Transfer of a Gloss Language Model for Semantic Change Detection // Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue. — 2021.
- [14] Rodina Julia, Kutuzov Andrey. RuSemShift: a dataset of historical lexical semantic change in Russian // Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona, Spain (Online) : International Committee on Computational Linguistics, 2020. — Dec. — P. 1037–1047. — Access mode: <https://www.aclweb.org/anthology/2020.coling-main.90>.
- [15] Ryzhova Anastasiia, Ryzhova Daria, Sochenkov Ilya. Detection of Semantic Changes in Russian Nouns with Distributional Models and Grammatical Features // Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue. — 2021.
- [16] Schlechtweg Dominik, Schulte im Walde Sabine, Eckmann Stefanie. Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). — New Orleans, Louisiana : Association for Computational Linguistics, 2018. — Jun. — P. 169–174. — Access mode: <https://www.aclweb.org/anthology/N18-2027>.
- [17] SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection / Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen et al. // Proceedings of the Fourteenth Workshop on Semantic Evaluation. — Barcelona (online) : International Committee for Computational Linguistics, 2020. — Dec. — P. 1–23. — Access mode: <https://www.aclweb.org/anthology/2020.semeval-1.1>.
- [18] Tang Xuri. A state-of-the-art of semantic change computation // Natural Language Engineering. — 2018. — Vol. 24, no. 5. — P. 649–676.
- [19] Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change / Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Florence, Italy : Association for Computational Linguistics, 2019.
- [20] Tracing cultural diachronic semantic shifts in Russian using word embeddings: test sets and baselines / Vadim Fomin, Daria Bakshandaeva, Julia Rodina, Andrey Kutuzov // Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue. — 2019. — P. 203–218. — Access mode: <http://www.dialog-21.ru/media/4598/fominvplusetal-116.pdf>.
- [21] Unsupervised Cross-lingual Representation Learning at Scale / Alexis Conneau, Kartikay Khandelwal, Naman Goyal et al. // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — Online : Association for Computational Linguistics, 2020. — Jul. — P. 8440–8451. — Access mode: <https://www.aclweb.org/anthology/2020.acl-main.747>.



- [22] A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains / Dominik Schlechtweg, Anna Häty, Marco Del Tredici, Sabine Schulte im Walde // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Florence, Italy : Association for Computational Linguistics, 2019. — Jul. — P. 732–746. — Access mode: <https://www.aclweb.org/anthology/P19-1072>.

A RuShiftEval gold datasets

1. 1-2: change from the pre-Soviet to Soviet times;
2. 2-3: change from the Soviet to the post-Soviet times;
3. 1-3: change from the pre-Soviet to the post-Soviet times.

DEVELOPMENT SET					
WORD	TRANSLITERATION	TRANSLATION	CHANGE SCORE		
			1-2	2-3	1-3
верховье	verhovje	upper reaches	3.68	3.74	3.87
возраст	vozrast	age	3.47	3.69	3.58
завод	zavod	factory/breeding farm	3.22	3.65	3.52
закладка	zakladka	foundation/bookmark/hidden artifact	1.93	1.74	1.74
земля	zemlja	earth/land/soil	2.83	2.8	2.28
лох	loh	salmon/silver-berry/easy victim	1.07	2.94	1.04
помощник	pomoštšnik	assistant	3.38	3.56	3.28
пролетарий	proletarij	proletarian	3.4	3.58	3.44
промышленность	promyšlennost'	industry	3.24	3.51	3.47
радикал	radikal	radical	1.42	1.68	2.01
спутник	sputnik	fellow traveler/satellite/sputnik	2.96	1.81	1.94
четверть	tšetvert	quarter	2.25	2.96	3.07

TEST SET					
WORD	TRANSLITERATION	TRANSLATION	CHANGE SCORE		
			1-2	2-3	1-3
авторитет	avtoritet	authority/prestige	3.23	2.95	2.84
амбиция	ambitsia	ambition	3.11	3.44	3.33
апостол	apostol	apostle/disciple	3.49	3.42	3.42
благодарность	blagodarnost'	gratitude/appreciation/thankfulness	3.23	3.56	3.65
блин	blin	pancake/damn	3.21	1.66	2.57
блондин	blondin	blonde (male)	3.94	3.92	3.95
брат	brat	brother	3.22	3.01	3.27
бригада	brigada	brigade/gang/team	2.8	2.71	3.08
веер	vejer	fan	2.55	2.43	2.44
век	vek	century/age	3.2	3.21	2.98
вызов	vyzov	call/challenge/summons	2.17	2.1	2.03
головка	golovka	(small) head	2.20	1.67	2.19
грех	greh	sin/fault	3.48	2.98	2.92
дух	duh	spirit/ghost/scent	2.32	1.63	1.88
дядька	djadka	uncle/man/(male) tutor	2.59	3.03	2.68
дядя	djadja	uncle/man	3.37	3.39	3.29
железо	železo	iron	2.2	2.56	2.40
жест	žest	tin/horror	3.23	3.38	3.41
живот	život	stomach/belly/life	2.91	3.44	2.76
заблуждение	zabluždenije	delusion	3.5	3.62	3.55
издательство	izdatelstvo	publishing house	3.53	3.86	3.45
итальянец	italjanets	Italian	3.70	3.6	3.67
кабан	kaban	boar	3.6	3.32	3.30
карман	karman	pocket	3.46	3.47	3.56
крушение	krušenje	collapse	2.75	2.78	2.6
крыша	kryša	roof	3.57	3.0	2.82
кулиса	kulisa	wings	3.16	3.17	3.24
лечение	letsenije	cure	3.65	3.74	3.68
линейка	lineika	carriage/ruler/series of goods	1.87	1.37	1.22
лишение	lišenije	deprivation	2.94	2.07	2.33
локоть	lokot	elbow	3.27	3.41	3.73
любовник	ljubovnik	lover	3.45	3.71	3.65
любовь	ljubov	love	3.29	2.97	3.07
маньяк	manjak	maniac	3.08	3.01	3.11
монстр	monstr	monster	2.6	2.38	2.04
наволочка	navolotška	pillowcase	3.61	3.83	3.92
название	nazvanije	name/title	3.48	3.48	3.43

WORD	TRANSLITERATION	TRANSLATION	CHANGE SCORE		
			1-2	2-3	1-3
наложение	naloženije	imposition	1.95	2.06	1.78
облако	oblako	cloud	3.17	3.0	3.16
обоснование	obosnovanije	grounds	3.74	3.5	3.58
огонь	ogon	fire	2.10	2.13	2.46
памятник	pamjatnik	monument	2.88	2.83	2.82
пафос	pafos	pathos	3.34	3.27	3.41
писк	pisk	squeak	3.21	3.0	2.53
план	plan	plan	2.67	2.27	2.54
поколение	pokolenie	generation	3.43	3.58	2.8
половинка	polovinka	half	2.51	2.75	2.62
полоса	polosa	stripe/ribbon/lane/runway	1.83	1.5	1.41
полость	polost	cavity/foot hide	2.23	1.88	2.56
полукруг	polukrug	semicircle	2.78	3.13	3.08
понедельник	ponedelnik	Monday	3.77	3.86	3.86
поставщик	postavštšik	supplier	3.56	3.44	3.25
поэзия	poezia	poetry	3.22	3.66	3.56
правда	pravda	truth/reality	3.13	2.94	2.96
предательство	predatelstvo	betrayal	3.67	3.48	3.8
прецедент	pretsedent	precedent	3.52	3.8	3.53
проникновение	proniknovenije	penetration	2.75	2.68	2.53
прорыв	proryv	breakthrough	2.08	2.05	2.05
путь	put'	way	2.41	2.04	2.3
размышление	razmyšlenije	reflection	3.52	3.55	3.62
ранец	ranets	backpack	3.6	3.53	3.38
расчет	rastšot	calculation/settlement	2.0	1.95	2.0
риторика	ritorika	rhetoric	3.06	2.95	2.93
роспись	rospis	mural/signature/list	1.43	2.98	1.57
сверстник	sverstnik	age-mate	3.86	3.86	3.82
связка	svjazka	ligament/vocal cords/mutual connection	2.33	1.96	1.77
собрat	sobrat	fellow	3.45	3.32	3.32
совершенство	soveršenstvo	perfection	2.95	3.16	3.08
советчик	sovettik	adviser	3.22	3.48	3.42
союзник	sojyznik	ally	3.66	3.47	3.75
список	spisok	list	3.28	3.31	3.05
ссылка	ssylka	exile/link	2.87	2.04	1.93
стена	stena	wall	3.1	3.16	3.32
стипендия	stipendia	scholarship	3.8	3.71	3.56

WORD	TRANSLITERATION	TRANSLATION	CHANGE SCORE		
			1-2	2-3	1-3
стол	stol	table/diet	3.50	3.16	3.25
тачка	tachka	wheelbarrow/car	3.39	1.94	1.89
тупик	tupik	deadlock	3.17	2.83	3.14
увольнение	uvolnenie	furlough/layoff	3.21	3.53	3.32
углеводород	uglevodorod	hydrocarbon	3.68	3.31	3.2
удобство	udobstvo	convenience	2.43	2.42	2.51
уклад	uklad	setup	3.33	3.42	3.42
университет	universitet	university	3.54	3.7	3.72
установление	ustanovlenie	establishment	2.28	2.26	2.40
фаворит	favorit	favorite	3.15	2.53	2.84
формат	format	format	2.84	2.02	1.81
формула	formula	formula	2.81	2.26	2.57
хозяйка	hozjaika	hostess	3.25	3.22	3.42
хор	hor	choir	2.66	2.87	2.22
хрен	hren	horseradish/dick/old fart	1.8	2.26	1.6
цензура	tsenzura	censorship	3.49	3.46	3.45
центр	tsentr	center	2.14	1.83	1.87
цифра	tsifra	digit/number	2.96	2.87	3.19
частица	tšastitsa	part/particle	1.96	2.33	2.2
чек	tšek	check	2.37	1.95	2.65
штаб	štab	headquarters	3.63	3.38	3.5
эшелон	ešelon	echelon	2.92	2.28	2.33
юбилей	jubilei	anniversary/jubilee	3.68	3.7	3.78
ядро	jadro	cannonball/core/nucleus	1.55	1.91	1.47
ясли	jasli	nursery/manger	2.28	3.0	1.9

Appendix C: Three-part diachronic semantic change dataset for Russian

Three-part diachronic semantic change dataset for Russian

Andrey Kutuzov

University of Oslo
Norway

andreku@ifi.uio.no

Lidia Pivovarova

University of Helsinki
Finland

lidia.pivovarova@helsinki.fi

Abstract

We present a manually annotated lexical semantic change dataset for Russian: *RuShiftEval*. Its novelty is ensured by a single set of target words annotated for their diachronic semantic shifts across three time periods, while the previous work either used only two time periods, or different sets of target words. The paper describes the composition and annotation procedure for the dataset. In addition, it is shown how the ternary nature of *RuShiftEval* allows to trace specific diachronic trajectories: ‘changed at a particular time period and stable afterwards’ or ‘was changing throughout all time periods’. Based on the analysis of the submissions to the recent shared task on semantic change detection for Russian, we argue that correctly identifying such trajectories can be an interesting sub-task itself.

1 Introduction

This paper describes *RuShiftEval*: a new dataset of diachronic semantic changes for Russian words. Its novelty in comparison with prior work is its multi-period nature. Until now, semantic change detection datasets focused on shifts occurring between **two** time periods. On the other hand, *RuShiftEval* provides human-annotated degrees of semantic change for a set of Russian nouns over **three** time periods: pre-Soviet (1700-1916), Soviet (1918-1990) and post-Soviet (1992-2016). Notably, it also contains ‘skipping’ comparisons of pre-Soviet meanings versus post-Soviet meanings. Together, this forms three subsets: *RuShiftEval-1* (pre-Soviet VS Soviet), *RuShiftEval-2* (Soviet VS post-Soviet) and *RuShiftEval-3* (pre-Soviet VS post-Soviet).

The three periods naturally stem from the Russian history: they were radically different in terms of life realities and writing and practices, which is reflected in the language. As an example, the word *дядька* lost its ‘tutor of a kid in a rich family’

sense in the Soviet times, with only the generic ‘adult man’ sense remaining. Certainly, language development never stops and Russian also gradually evolved within those periods as well, not only on their boundaries. However, in order to create a usable semantic change dataset, one has to draw the boundaries somewhere, and it is difficult to come up with more fitting ‘changing points’ for Russian.

RuShiftEval can be used for testing the ability of semantic change detection systems to trace long-term multi-point dynamics of diachronic semantic shifts, rather than singular change values measured by comparing two time periods. As such, *RuShiftEval* was successfully employed in a recent shared task on semantic change detection for Russian (Kutuzov and Pivovarova, 2021).

2 Related work

Automatic detection of word meaning change is a fast growing research area (Kutuzov et al., 2018; Tahmasebi et al., 2018). Evaluation of this task is especially challenging; *inter alia*, it requires gold standard annotation covering multiple word usages.

The common practice is to annotate pairs of sentences as using a target word in either the same or different senses. It was introduced for the word sense disambiguation task in (Erk et al., 2013), while (Schlechtweg et al., 2018) proposed methods to aggregate pairwise annotations for semantic change modeling; one of them, the COMPARE metrics, is used in *RuShiftEval*.

A similar approach was used for the SemEval’20 shared task on semantic change detection (Schlechtweg et al., 2020): annotators labeled pairs of sentences, where some pairs belonged to the same periods and some to different ones. This annotation resulted in a diachronic word usage graph, which was then clustered to obtain sepa-

rate word senses and their distributions between time periods (Schlechtweg et al., 2021).

The pairwise sentence annotation has been used in creating another semantic change dataset for Russian, *RuSemShift* (Rodina and Kutuzov, 2020). We use the same annotation procedure and rely on the same corpus, i.e. Russian National Corpus (RNC) split into pre-Soviet, Soviet and post-Soviet sub-corpora. However, *RuSemShift* features two sets of words: one for the changes between the pre-Soviet and Soviet periods, and another for the Soviet and post-Soviet periods. The new *RuShiftEval* dataset, which we present in this paper, uses a *joint word set* allowing for tracing each word across three time periods. In addition, we directly annotate semantic change between the pre-Soviet and post-Soviet periods, skipping the Soviet one.

3 Dataset Construction

3.1 Word List Creation

In building the dataset, we relied on the graded view on word meaning change (Schlechtweg et al., 2021): for each word in the dataset, we measure a *degree of change* between pairs of periods, rather than making a binary decision on whether its sense inventory changed over time. The measure relies on pairwise sentence annotations, where each pair of sentences is processed by at least three annotators.

Compiling the target-word set, we needed to ensure two main conditions: (i) the dataset contains many ‘interesting’ words, i.e. words that changed their meaning between either pair of periods; (ii) not all words in the dataset actually changed their meaning. We followed the same procedure as in (Kutuzov and Kuzmenko, 2018; Rodina and Kutuzov, 2020; Schlechtweg et al., 2020): first, select changing words, and then augment them with *fillers*, i.e. random words following similar frequency distribution across three time periods.

Technically, it was possible to populate the target word set automatically, using any pre-trained language model (LM) for Russian and some measure of distance between word representations in different corpora. However, we wanted our target words choice to be motivated linguistically rather than influenced by any LM architecture. Therefore, to find changing words, we first consulted several dictionaries of outdated or, on the contrary, the most recent Russian words, such as (Novikov, 2016; Basko and Andreeva, 2011; Skljarevsky, 1998). Unfortunately, dictionaries provided less examples than we

needed: they often contain archaisms, neologisms, multi-word expressions, and words which are infrequent in the corpus or not used in the meanings specified in the dictionaries.

However, we discovered that some changing words could be found in papers on specific linguistics problems. For example, the word облако (‘cloud’) was found in a paper on the Internet language (Baldanova and Stepanova, 2016); стол (‘table/diet’)—in an article discussing the language of one story by Pushkin (M., 2016). Finally, to find some of the target words, we used our intuition as educated native speakers. Out of 50 words, 13 were found in dictionaries, 10 invented by ourselves and the rest 27 found in articles on more specific topics. Regardless the initial word origin, we manually checked that all words occur at least 50 times in each of the three sub-corpora and that the distinctive sense is used several times.

Fillers (selected for each target word) are sampled so that they belong to the same part of speech—nouns in our case—and their frequency percentile is the same as the target word frequency percentile in all three periods. The aim here is to ensure that frequency cannot be used to distinguish the target words from fillers.¹ For *RuShiftEval*, we sampled two filler words for each target word.

The final dataset consists of 111 Russian nouns, where 12 words form a development set and 99 words serve as a test set. Since the annotation procedure is the same as for *RuSemShift* (Rodina and Kutuzov, 2020), one can use one of these resources as a training set and then evaluate on another.

3.2 Annotation

Annotators’ guidelines were identical to those in *RuSemShift* (Rodina and Kutuzov, 2020). To generate annotation tasks, we sampled 30 sentences from each sub-corpus and created sentence pairs. We ran this sampling independently for all three period pairs. The sentences were accompanied by one preceding and one following sentence, to ease the annotators’ work in case of doubt. The task was formulated as labeling on a 1-4 scale, where 1 means the senses of the target word in two sentences are unrelated, 2 stands for ‘distantly related’, 3 stands for ‘closely related’, and 4 stands for ‘senses are identical’ (Hätty et al., 2019). Annotators were also allowed to use the 0 (‘cannot decide’) judgments.

¹Indeed, there is no significant correlation between frequency differences and the aggregated relatedness scores from our gold annotation.

Time bins	α	ρ	JUD	0-JUD
Test set (99 words)				
RuShiftEval-1	0.506	0.521	8 863	42
RuShiftEval-2	0.549	0.559	8 879	25
RuShiftEval-3	0.544	0.556	8 876	31
Development set (12 words)				
RuShiftEval-1	0.592	0.613	1 013	7
RuShiftEval-2	0.609	0.627	1 014	3
RuShiftEval-3	0.597	0.632	1 015	2

Table 1: *RuShiftEval* statistics. α and ρ are inter-rater agreement scores as calculated by Krippendorff’s α (ordinal scale) and mean pairwise Spearman ρ . JUD is total number of judgments and 0-JUD is the number of 0-judgments (‘cannot decide’).

They were excluded from the final datasets, but their number was negligible anyway: about 100 out of total 30 000.

The annotation was carried out on the Yandex.Toloka crowd-sourcing platform.² We employed native speakers of Russian, older than 30, with a university degree. To ensure the annotation quality, the authors themselves annotated about 20 control examples for each pair of periods. We chose the most obvious cases of 1 and 4 for this; annotators who answered incorrectly (not with the exactly matching grade), were banned from the task for 24 hours. The inter-rater agreement statistics and the number of judgments in each *RuShiftEval* subset are shown in Table 1. The agreement is on par with other semantic change annotation efforts: (Schlechtweg et al., 2020) report Spearman correlations ranging from 0.58 to 0.69, (Rodina and Kutuzov, 2020) report Krippendorff’s α ranging from 0.51 to 0.53.³ Each subset was annotated by about 100 human raters, more or less uniformly ‘spread’ across annotation instances, with the only constraint being that each instance must be annotated by three different persons.

Finally, the degrees of semantic change for each word between a pair of periods were calculated using the COMPARE metrics (Schlechtweg et al., 2018), which is the average of pairwise relatedness scores. Interestingly, some words initially sampled as fillers—e.g. ядро (‘cannonball or

core/nucleus’)—ended up among most changed according to the annotation. Also some words from the initial set were annotated as relatively stable. This happened because the distinctive sense was rare or because annotators’ opinion diverged from linguistic knowledge in the dictionaries. For example, for the word бригада (‘brigade/gang/team’) dictionaries list two distinct senses—a military and a civil one. However, in most cases the annotators considered these senses identical or closely related.

The dataset is publicly available, including the raw scores assigned by annotators.⁴

4 Diachronic trajectory types

RuShiftEval allows tracing multi-hop dynamics of semantic change. A similar analysis of diachronic word embedding series or ‘trajectories’ was conducted in (Kulkarni et al., 2015) and (Hamilton et al., 2016b), but the former focused on change point detection, and the latter on finding general laws of semantic change. With manually annotated *RuShiftEval* dataset we were able to move further and identify at least three different types of changing trajectories: 1) changes in every period pair; 2) change in the Soviet period as compared to the pre-Soviet period; 3) change in the post-Soviet period as compared to the Soviet period.

Since approximately a half of the words in the dataset did not change their meaning they exhibit a fourth, trivial type of trajectory, where all three distances are small. In principle there could be a fifth type of trajectory, where difference between pre-Soviet and post-Soviet periods is substantially smaller than between other period pairs, which would mean that a word was used in a new sense during the Soviet time but then came back to its original meaning. However, we did not find any words following this trajectory type and not sure whether this behavior is theoretically plausible.

Table 2 shows examples of nouns belonging to three non-stable trajectory types. Below we explain the semantic change processes for them.

1. The word закладка belongs to the type 1. Its dominant sense in the pre-Soviet period was ‘foundation’ (as in ‘*The foundation of the new church building took place yesterday*’). In the Soviet times, the ‘bookmark’ sense emerged (it was already present before, but very rare). Then, the post-Soviet time period saw the emergence of two

²<https://toloka.yandex.ru/>

³Note it does not make much sense to report correlations for individual annotators (‘data columns’), since in our crowd-working setup, the columns are not associated with particular persons.

⁴https://github.com/akutuzov/rushifteval_public

Type	Examples	Baseline	Top
1	закладка ('foundation/bookmark/hidden artifact'), линейка ('carriage/ruler/series of goods'), центр ('center')	0.5	1.0
2	дядька ('tutor/adult man'), живот ('life/belly/stomach'), лох ('salmon/silver-berry/easy victim, stupid person'), роспись ('list/painting'), ядро ('cannonball/core/nucleus')	1.0	1.0
3	полоса ('stripe/ribbon/lane/runway'), связка ('ligament/vocal cords/mutual connection'), спутник ('fellow traveler/satellite/sputnik'), ссылка ('exile/link'), тачка ('wheelbarrow/car'), формат ('format')	0.4	0.8-1.0

Table 2: Semantic change trajectory types in *RuShiftEval* and the percentage of words with correctly captured type for the baseline and the 4 best shared task submissions (see 4.1).

new senses, both through widening processes: ‘tab’ (in graphical user interfaces) and ‘booby-trapping’ or ‘something hidden’ (often about illegal drugs cached by a distributor). Thus, low relatedness scores are observed across all possible pairs: the word is used differently in each time period.

2. The word *ядро* can mean either ‘cannonball’ or ‘core/kernel/nucleus’. It belongs to the type 2. In the Soviet period, the first sense almost disappeared (because artillery stopped using cannonballs in the 20th century), while the latter sense became more frequent. After this reduction, the meaning was stable, with no changes in the post-Soviet period.

3. The word *тачка* (‘wheelbarrow’) belongs to the type 3. It was stable until the end of the Soviet period, but in the post-Soviet times, *тачка* acquired a new colloquial sense of ‘car’, quite common even in written texts. This led to divergence from both Soviet and pre-Soviet periods.

Semantic trajectory types could be visualized as time relatedness graphs; see Figure 1. Nodes of the graph are time periods, and edge widths represent the COMPARE score (see 3.2) for each pair of periods.⁵ Thus, thicker edges denote stable meaning, while thinner and more transparent edges show a change. Each trajectory type has its own characteristic pattern of edge widths. For example, in the graph for *тачка* (the rightmost plot), the edges connecting the post-Soviet node to two other nodes are much thinner than the edge between the pre-Soviet and post-Soviet nodes. This signals a change in the post-Soviet times (trajectory type 3).

⁵Note that in most cases it is impossible to use nodes relative positions on the plot to reflect relatedness scores: one can’t change the length of an edge in a triangle without also changing the length of at least one other edge.

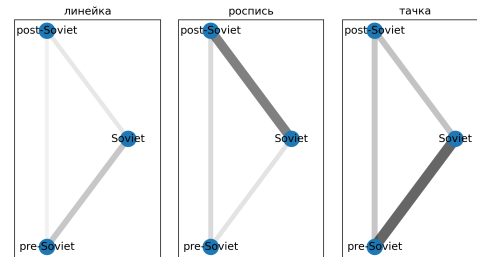


Figure 1: Time relatedness graphs for words belonging to different semantic trajectory types (from left to right): *линейка* (‘carriage/ruler/series of goods’) (1), *роспись* (‘list/painting’) (2), *тачка* (‘wheelbarrow/car’) (3).

Note that the annotation process and the definition of the COMPARE score itself do not guarantee perfect capturing of semantic changes. One example—made clear by the multi-period nature of *RuShiftEval* design—is the word *радикал* (‘radical’). Its relatedness scores are low across all time period pairs, suggesting that it experienced sequential changes similar to *закладка*. However, in fact, throughout all the times covered by *RuShiftEval*, this word had the same two persistent senses: political and chemical. Since their probabilities were almost equal, many randomly sampled sentence pairs contained the word *радикал* in two different senses, which led to low COMPARE scores. In this case, it stems from strong and persistent ambiguity of the word, not from diachronic semantic change. This limitation of the COMPARE metrics was already described in (Schlechtweg et al., 2018).

Another potential flaw is sampling variability. For annotation, we sampled 30 sentences with a target word from each time period for each comparison. Since our relatedness graph has three edges,

each word is represented with two samples. As it turned out, in some cases different samples can yield quite different picture of sense distributions.

Let us manually analyze the word *полость* ('cavity/hide to cover one's legs in an open cart'). Since horse-driven carts disappeared just a few years after the beginning of the Soviet period, one might expect the second sense to be lost in Soviet times and never to appear again. However, the relatedness between the Soviet and post-Soviet time periods (1.9) is even lower than between the pre-Soviet and Soviet periods (2.2), as if the word experienced another semantic shift. In fact, it is a random sampling artifact. In the 30 sentences from the Soviet period sampled for the 'pre-Soviet:Soviet' pair, only 4 used *полость* in this archaic sense. But in the 30 sentences from the same period sampled for the 'Soviet:post-Soviet' pair, this number grew to 10, 2.5 times more (mostly in fiction texts, where the plot is set in the pre-Soviet times). As a result, the Soviet usage pattern looks like it is different from the post-Soviet one, although in fact no shift has happened (as evident both from linguistic intuition of Russian speakers and from the Fisher exact test which in this case yields $p = 0.13$). The frequency of *полость* in the Soviet sub-corpus is about 600, so both samples together cover only 10% of the full concordance. Without manually annotating all six hundred occurrences, it is difficult to tell which sample is more representative of the real word usage in the Soviet times. It would be better to increase the sample size as much as possible: 30 is arguably already on the border.

4.1 Trajectory detection task?

The *RuShiftEval* dataset was used to evaluate the systems participating in a shared task on lexical semantic change detection for Russian (Kutuzov and Pivovarova, 2021). How good these submissions are in capturing the trajectory types described in the previous section? In this subsection, we describe a toy experiment to address this question.

For simplicity, we will use only 11 example words from Table 2 which appear in the *RuShiftEval* evaluation set (this excludes *закладка*, *лох* and *спутник*, since they appear in the development set only). Then a set of criteria is established for the system predictions, corresponding to each of the three trajectory types. We consider a system successful in capturing a word with the **trajectory 2** if the predicted relatedness score is higher for the

'Soviet:post-Soviet' pair than for other two pairs. For the words with the **trajectory 3**, the relatedness score for the 'pre-Soviet:Soviet' pair must be the highest among all pairs. For the words with the **trajectory 1**, the percentile ranks of the relatedness scores for all three sub-sets must be below 50 (admittedly, this is an *ad hoc* criterion, but it is used here just to give an example of how the task can be set up). Thus, at least for the trajectory types 2 and 3, this resembles a simple ranking task: not across target words within one period pair, but for one target word across three period pairs. At the same time, the trajectory type 1 (changes in every period) does not quite fit into this frame.

We compared the baseline system (which used static diachronic word embeddings and the local neighbors method from (Hamilton et al., 2016a)) and four best systems (employing contextualized language models: ELMo, BERT or XLM-R). The results are presented in Table 2. All of the best submissions captured the **trajectory 1** for all two target words, but the baseline method failed for *центр* (its percentile rank in *RuShiftEval-1* is more than 60). For the **trajectory 3**, the top systems are considerably better than the baseline method. For example, according to the baseline method, *полоса* experienced its strongest change in the Soviet times, while in fact it was in the post-Soviet period. Only for the **trajectory 2**, the baseline is on par with the winners of the shared task.

This analysis is rather preliminary, but it shows that the systems performance in correctly detecting diachronic trajectories does to some extent correlate with their performance in the 'traditional' semantic change ranking (with binary datasets, like in the SemEval 2020 Shared Task 1). We believe that this can be an interesting sub-task within the larger field of semantic change detection, once more datasets like *RuShiftEval* are available and more formal definitions of 'capturing the trajectory successfully' are developed.

5 Conclusion

We presented *RuShiftEval*, a novel dataset of diachronic semantic changes in Russian nouns across three time periods, using the same set of target words for all comparisons. We also conducted a preliminary analysis of how *RuShiftEval* can be used in tracing diachronic semantic trajectories, and how current change detection systems for Russian deal with this potentially interesting task.

Acknowledgments

The annotation effort for *RuShiftEval* was supported by the Russian Science Foundation grant 20-18-00206. This work has been partially supported by the European Union Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDDIA).

References

- Marina Baldanova and Irina Stepanova. 2016. Metaforizatsiya kak put' razvitiya semanticheskikh neologizmov v yazyke interneta (metaphorization as a way of developing semantic neologisms in the language of the internet). In *Russian*.
- Nina Basko and Irina Andreeva. 2011. *Slovar' ustarevshey leksiki k proizvedeniyam russkoy klasiki* (Dictionary of obsolete vocabulary for the works of Russian classics). In Russian.
- Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. *Measuring word meaning in context*. *Computational Linguistics*, 39(3):511–554.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. *Cultural shift or linguistic drift? comparing two computational measures of semantic change*. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. *Diachronic word embeddings reveal statistical laws of semantic change*. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.
- Anna Hättö, Dominik Schlechtweg, and Sabine Schulte im Walde. 2019. *SURel: A gold standard for incorporating meaning shifts into term extraction*. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 1–8, Minneapolis, Minnesota. Association for Computational Linguistics.
- Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635, Florence, Italy.
- Andrey Kutuzov and Elizaveta Kuzmenko. 2018. Two centuries in two thousand words: neural embedding models in detecting diachronic lexical changes. *Quantitative Approaches to the Russian Language*, page 95.
- Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. *Diachronic word embeddings and semantic shifts: a survey*. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Andrey Kutuzov and Lidia Pivovarov. 2021. *RuShiftEval: a shared task on semantic shift detection for Russian*. In print.
- Elmi A. M. 2016. *Izmeneniya znacheniya odnoznachnykh imen sushchestvitel'nykh, upotreblennykh v povesti as pushkina "grobovshchik"* (changes in the meaning of unambiguous nouns used in as pushkin's story "the undertaker"). In *Russian*.
- Vladimir Novikov. 2016. *Dictionary of buzzwords. The linguistic picture of our time*. In Russian.
- Julia Rodina and Andrey Kutuzov. 2020. *RuSemShift: a dataset of historical lexical semantic change in Russian*. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1037–1047, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. *SemEval-2020 task 1: Unsupervised lexical semantic change detection*. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.
- Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. *DWUG: A large resource of diachronic word usage graphs in four languages*. *arXiv preprint arXiv:2104.08540*.
- Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. *Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change*. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.
- Skljarevsky, editor. 1998. *Tolkovyy slovar' russkogo yazyka kontsa XX veka. Yazykovyye izmeneniya*. (Explanatory dictionary of the Russian language at the end of the XX century. Language changes). In Russian.
- Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. *Survey of computational approaches to diachronic conceptual change*. *arXiv preprint arXiv:1811.06278*.

A Transliterations of Russian words mentioned in the article

WORD	TRANSLITERATION	TRANSLATION
бригада	brigada	brigade/gang/team
дядька	djadka	uncle/man/(male) tutor
живот	život	stomach/belly/life
закладка	zakladka	foundation/bookmark/hidden artifact
линейка	lineika	carriage/ruler/series of goods
лох	loh	salmon/silver-berry/easy victim
облако	oblako	cloud
полоса	polosa	tripe/ribbon/lane/runway
полость	polost	cavity/foot hide
радикал	radikal	radical
ропись	rospis	mural/signature/list
связка	svjazka	ligament/vocal cords/mutual connection
спутник	sputnik	fellow traveler/satellite/sputnik
ссылка	ssylka	exile/link
стол	stol	table/diet
тачка	tachka	wheelbarrow/car
формат	format	format
центр	tsentr	center
ядро	jadro	cannonball/core/nucleus

Appendix D: Multilingual Topic Labelling of News Topics using Ontological Mapping

Multilingual Topic Labelling of News Topics using Ontological Mapping

Elaine Zosa^[0000–0003–2482–0663], Lidia Pivovarov^[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.

2 E. Zosa, et al.

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 $X = (x_1, \dots, x_n)$ 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>

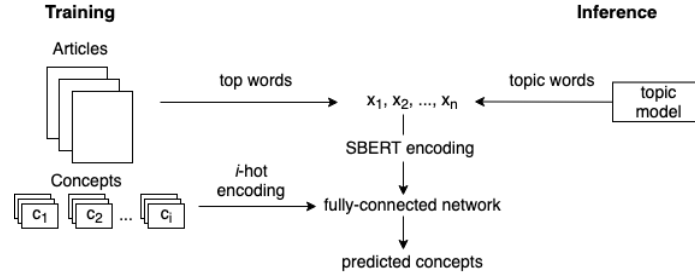


Fig. 1. News concepts prediction pipeline.

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, \dots, 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 HuggingFace³ for 5 epochs. We use the AdamW optimizer with weight decay set to 0.01. We refer to this as the **mbart** model⁴. 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].

4 E. Zosa, et al.

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 $X = (x_1, \dots, x_n)$ 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 topic⁶. 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			
<i>baseline: top 5 terms</i>	<i>89.47</i>	<i>88.08</i>	<i>88.49</i>
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			
<i>baseline: top 5 terms</i>	98.17	96.58	97.32
ontology-tfidf	97.00	95.25	96.04
ontology-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

6 E. Zosa, et al.

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 (<i>to drive</i>), hamilton, mercedes
Gold	formula, formulat, formula 1, f1, formula-auto, aika-ajot (<i>time trial</i>), moottoriurheilu (<i>motor sport</i>)
rnn-tfidf	autourheilu (<i>auto sport</i>), urheilutapahtumat (<i>sports event</i>), mm-kisat (<i>world championship</i>), urheilu (<i>sport</i>), urheilijat (<i>athletes</i>)
transformer-sent	urheilutapahtumat (<i>sports event</i>), mm-kisat (<i>world championship</i>), urheilu (<i>sport</i>), autourheilu (<i>auto sport</i>), kansainväliset (<i>international</i>)
mbart-sent	autourheilu moottoriurheilu, urheilutapahtumat, mm-kisat , urheilijat pelaajat, urheilu
ontology-sent	ID: 15000000, fi: <u>urheilu</u> , en: sport; ID: 15039000, fi: <u>autourheilu moottoriurheilu</u> , en: motor racing; ID: 15073000, fi: <u>urheilutapahtumat</u> , en: sports event; ID: 15039001, fi: <u>formula 1</u> , en: formula one; ID: 15073026, fi: <u>mm-kisat</u> , 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: <u>film festival</u> , fi: elokuvajuhlat; ID: 04010003, en: cinema industry, fi: elokuvateollisuus; ID: 08000000, en: <u>human interest</u> , fi: human interest; ID: 01022000, en: culture (general), fi: kulttuuri yleistä; ID: 04010000, en: <u>media</u> , fi: medialaous

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)

8 E. Zosa, et al.

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., Pivovarov, 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, L., 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)

Appendix E: Elastic Embedded Background Linking for News Articles with Keywords, Entities and Events

Elastic Embedded Background Linking for News Articles with Keywords, Entities and Events

Luis Adrián Cabrera-Diego*
University of La Rochelle, L3i
La Rochelle, France
a.cabrera@jsumundi.com

Emanuela Boros
University of La Rochelle, L3i
La Rochelle, France
emanuela.boros@univ-lr.fr

Antoine Doucet
University of La Rochelle, L3i
La Rochelle, France
antoine.doucet@univ-lr.fr

ABSTRACT

In this paper, we present a collection of five flexible background linking models created for the News Track in TREC 2021 that generate ranked lists of articles to provide contextual information. The collection is based on the use of sentence embeddings indexes, created with Sentence BERT and Open Distro for Elasticsearch. For each model, we explore additional tools, from keywords extraction using YAKE, to entity and event detection, while passing through a linear combination. The associated code is available online as open-source software.

CCS CONCEPTS

• Information systems → Retrieval models and ranking; Language models; Rank aggregation.

1 INTRODUCTION

With the massification of the internet and mobile devices, such as smartphones, people have started to access news more frequently from digital sources than printed ones [11, 13]. This has meant that newspaper publishers have had to focus more on the digital experience and perform users' behavioral analysis for providing tools such as news recommendation [33]. Furthermore, as Pranjić et al. [27] indicate, linking news to other relevant articles can improve businesses' websites metrics such as user engagement and average time on page. Subsequently, this can improve revenues from ads or sponsored articles.

Therefore, in 2018 the *Text REtrieval Conference (TREC)* along with *The Washington Post*¹, decided to propose the News Track [30], a track where the goal is to enhance users' experience while reading news articles.

Since TREC 2020, the News Track is organized into two subtasks, *Background Linking* and *Wikification*. The former has been defined as the task where "given a news article, a system should retrieve other news articles that provide important context and/or background information that helps the reader better understand the query article" [29]. The latter exploits, as a means of contextualization, the linking of textual elements, such as concepts and artifacts, to an external knowledge-base, in this case to Wikipedia [31].

In this paper, we present the participation of the *L3i Laboratory* of the University of La Rochelle at the 2021 TREC News Track

Background Linking task. Our participation consisted of five different approaches that used, for instance, keyword extraction, entities, and events detection, but also sentence embeddings and linear combination.

2 RELATED WORK

Before TREC 2018 News Track, there is a reduced number of works that explore the use of news articles as a way to contextualize elements such as comments [1], tweets [14], or events [25].

Since the proposal of the News Track in TREC 2018, we have seen an increment of publications related to the contextualization of news articles using other news articles. Most of them are works explaining TREC participant systems [17, 20, 24]. However, we can find as well some other related outputs and analysis [12, 18].

More recently and besides TREC-related outputs, we can name the work of Pranjić et al. [27], where the authors explore different models to link background and related news articles in a Croatian corpus. Furthermore, Koloski et al. [19] explore the linking of cross-border news articles in Latvian and Estonian. Also, we can name the MIND dataset [33], a collection of news articles from *Microsoft News* that are associated with human behaviors, in order to explore news recommendation tasks.

3 DATA

For 2021, the TREC News Track organizers provided a corpus composed of 728,626 news articles and blog posts published by *The Washington Post* from January 2012 through December 2020. Each document, either news article or blog post, includes elements such as title, kicker (section header), body, author, images captions, and publication date. Also, TREC organizers delivered a list of 51 different topics, i.e. news articles, for which TREC News Track's participants had to propose background articles. For the 2021 edition of TREC News Track, the organizers also added a subtopic task, in which specific information, such as the background, is expected for each topic. In Figure 1, we present the topic structure used in the 2021 TREC News Track.

We first performed a pre-processing that consisted of parsing each document element, such as titles and captions, in order to get sentences. This pre-processing was done using *Turku Neural Parser* [16].

Once the documents were pre-processed, we decided to create embeddings for every document element using Sentence BERT [28], a fine-tuned BERT [10] which produces embeddings that can be compared using cosine similarity. Specifically, we made use of the

¹<https://www.washingtonpost.com/>

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
© 2021

```

<top>
<num>Number: xxx </num>
<docid>f30b7db4-cc51-11e6-a747-d03044780a02</docid>
<url>https://www.washingtonpost.com/local/public-safety/
homicides-remain-steady-in-the-washington-region/
2016/12/31/
f30b7db4-cc51-11e6-a747-d03044780a02_story.html</url>
<title>Topic title</title>
<desc>I would like to learn more about this topic</desc>
<narr>Traditional TREC narrative paragraph on the topic</narr>
<subtopics>
<sub num="1">This is the first subtopic.</sub>
<sub num="2">And this is the second one.</sub>
</subtopics>
</top>

```

Figure 1: Structure of TREC News Track 2021 topics, where the description and subtopics fields were added.

pre-trained model *stsb-mpnet-base-v2*² which at the time of the experiments was the most performing model available.

Due to limitations on how many tokens can be processed by Sentence BERT, i.e. 128 byte-pair encoding tokens, and to avoid losing vital information, we calculated the embeddings sentence by sentence. To be precise, we requested from Sentence BERT model the dense representation of each token in every sentence. The final dense representation of a text portion was obtained by averaging the dense vector of every token.

It should be indicated as well that we created composite vectors, in which we calculated the average embeddings based on multiple document elements: Title-Lead, Title-Body, and Title-Body-Captions. We also processed, in the same way, each topic provided by the TREC organizers, which notably included the creation of dense vectors for the narration or for the subtopics.

For retrieving documents from the corpus, we indexed the pre-processed data using *Open Distro for Elasticsearch*³ (ODFE), an *ElasticSearch*⁴ branch which implements a performing k-NN algorithm that can be used to retrieve documents using dense vectors, such as embeddings.⁵

In total, we indexed 728,500 articles from The Washington Post, which corresponded to 99.98% of the articles provided by TREC organizers. The code for pre-processing and indexing the data is publicly available in GitHub⁶. It should be noted, in the code, that the indexes contained more information than the one detailed in this work. However, not all the information was used for the creation of the submitted approaches.

²<https://huggingface.co/sentence-transformers/stsb-mpnet-base-v2>

³<https://opendistro.github.io>

⁴<https://www.elastic.co/>

⁵Although we use ODFE instead of ElasticSearch, the documentation of the latter is valid except for the dense vectors queries. Thus, we will point to ElasticSearch 7.12's documentation in specific cases.

⁶https://github.com/EMBEDIA/news_background_linking

4 EXPLORED APPROACHES

In this section, we describe in detail the five approaches we explored to provide background links for each topic:

- (1) **KWVec**: keywords and dense vectors to retrieve the related background articles;
- (2) **Lambda**: linear combination of multiple queries;
- (3) **300K_ENT_PH**: the articles retrieved by KWVec are re-ranked with the utilization of entities and event mentions;
- (4) **300K_ENT_PH_DN**: the articles retrieved and re-ranked by 300K_ENT_PH are again sorted depending on the description and the narrative field;
- (5) **Lambda_narr**: the outcome produced by the Lambda approach is followed by re-ranking the recommended articles using the narrative field.

Each of the following sections detail the five approaches used to provide subtopic background linking. These five approaches consist of re-rankings of the former approaches.

4.1 Run 1: KWVec

This approach consists of using keywords and dense vectors to retrieve the related background articles for a determined topic.

Specifically, we start by extracting unigram keywords from the text produced by the concatenation of the title, body, and captions.⁷ This is done using YAKE [9], an unsupervised keyword extractor. Once we have the unigram keywords, we obtain those related to the title by matching the title's unigrams and the obtained keywords.

The second step of KWVec consists of using a *boosting query*⁸, where a collection of queries are used to retrieve the documents, and another set is used to decrease their relevance.

To retrieve the documents, we submit three different queries to ODFE. Two of them ask ODFE to retrieve the documents that are relevant to the keywords found by YAKE. To be precise, we search title keywords in titles and body keywords in bodies. These queries are done through a *query string query*⁹.

Furthermore, as YAKE assigns a weight w to each keyword, we make use of these weights to increase or decrease the *query string query* relevance through the *boost* parameter. Nonetheless, as YAKE's weights interval is between $(0, \infty)$, where the lower the score the better, we modify it with Equation 1 to an interval of $(\inf, 0]$, where the higher the score the better.

$$KW_{weight} = \begin{cases} -\ln(w) & \text{if } w < 1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The third query retrieves the most relevant documents using ODFE's *exact k-NN* and cosine similarity.¹⁰ Specifically, the cosine similarity is calculated between the title-body dense vectors of the topic article and those found in the index.

It should be indicated that we modified ODFE's cosine similarity (s) score using Equation 2. The first reason is that ODFE's cosine

⁷We concatenate these text fields in order to get more relevant keywords. Focusing separately on smaller text portions, such as the title, produced less relevant keywords.

⁸<https://www.elastic.co/guide/en/elasticsearch/reference/7.12/query-dsl-boosting-query.html>

⁹<https://www.elastic.co/guide/en/elasticsearch/reference/7.12/query-dsl-query-string-query.html>

¹⁰<https://opendistro.github.io/for-elasticsearch-docs/docs/knn/knn-score-script/>

similarity is vertically translated, within the interval $[0, 2]$, to provide only positive scores. The second reason is to boost the cosine similarity by a scalar defined experimentally to 250 and prevent its fading with respect to the keywords scores.

$$Sim = \begin{cases} 250 \times (s - 1) & \text{if } s \geq 1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

We requested ODFE to reduce by 20% the relevance of documents that are associated with an unwanted kicker¹¹ and/or whose title was similar to the topic's. The former aspect was to reduce the relevance of articles that are frequently not used by The Washington Post's journalists. The latter aspect was calculated using *exact k-NN* and cosine similarity between titles dense vectors. We do this to avoid articles that might be considered relevant because they are either a duplicate of the topic article¹² or whose title is too similar.

4.2 Run 2: Lambda

Besides the previously described approach, we decided to explore a linear combination (see Equation 3) optimized through a Bayesian optimization algorithm [23]¹³. Through this optimization, our goal was to determine the weights (λ) that different queries scores (x), such as title similarity, should be given in order to achieve the highest nDCG evaluation. This approach is similar to the one used by Cabrera-Diego et al. [8] for merging different systems outputs.

$$\lambda_1 x_1 + \lambda_2 x_2 + \dots + \lambda_n x_n \quad (3)$$

For the Lambda approach, we explored four different independent queries¹⁴, *title to title*, *body to body*, *lead to title* and *lead to body*, using two methods, keywords and dense vectors. This gave a total of eight different independent queries used for the optimization. The queries based on keywords use the method presented in Section 4.1, while queries based on dense vectors used an unmodified version of ODFE's *exact k-NN* and cosine similarity. Furthermore, for the Lambda approach, we removed from the recommended articles those with an unwanted kicker (see Footnote 11).

To calculate the value of the different λ , we used as training data the sets provided by the organizers from previous years plus some additional articles that we annotated ourselves.¹⁵ Specifically, we requested ODFE to calculate¹⁶ the relevance score of the eight queries for each document for which we had a gold standard score. Then, the Bayesian optimization proposed different λ weights, in the interval of $[-10, 10]$, that optimized an objective function.

The objective function to be maximized by the Bayesian optimization is presented in Equation 4, where G is a weighted harmonic average, Q_1 and Q_3 are respectively the first and third quartile, and Q_2

is the median. These values are calculated based on the nDCG@10 scores obtained by each topic for all the years (2018-2020).¹⁷

$$G = \frac{5Q_1Q_2Q_3}{(Q_2Q_3) + (2.5Q_1Q_3) + (1.5Q_1Q_2)} \quad (4)$$

The weighted harmonic average presented in Equation 4 was defined to boost the median (Q_2) nDCG@10 score, but also to create a negatively skewed distribution of the nDCG@10 scores, by boosting the third quartile (Q_3). This would mean that we expect most of the nDCG@10 scores to have higher values rather than lower ones.

4.3 Run 3: 300K_ENT_PH

This approach extends the KWVec method with a re-ranking step applied after the relevant documents were retrieved by the ODFE query. Thus, since named entity recognition (NER) has been playing an important role in information seeking and retrieval, we propose to exploit knowledge about entities and their relationships (events) for re-evaluating the relevance of the query results. For this and for taking advantage of the annotation efforts from previous campaigns, we leverage the fine-grained entities defined by the organizers of the TAC KBP *Recognizing Ultra Fine-grained Entities* (RUFES) 2020¹⁸ and the events defined by the ACE 2005 evaluation campaign¹⁹.

4.3.1 Fine-grained Entities. The KBP 2020 RUFES dataset provided by the organizers consisted of the development source documents and evaluation source documents drawn from a collection of The Washington Post news articles. The development source corpus and the evaluation source corpus had approximately 100,000 articles each, from which 50 documents were annotated for the development set with entity types from an ontology that contains approximately 200 fine-grained entity types and that followed the same three-level x.y.z hierarchy as in the TAC-KBP 2019 EDL track [15]²⁰. For example, such an entity organized in a hierarchy is: *Photographer* is from an *Artist* that, in turn, is a subtype of *PER*²¹.

In order to benefit from the extraction of these entity types, we made use of our recently proposed model for coarse-grained and fine-grained named entity recognition [3–5, 7] that consists in a hierarchical, multitask learning approach, with a fine-tuned encoder based on BERT [10]. This model includes the use of a stack of Transformer [32] blocks on top of the BERT encoder. The multitask prediction layer consists of separate conditional random field (CRF) layers.

In Table 1, we explore two pre-trained and fine-tuned BERT *cased* models, BERT-base and BERT-large. We further consider the BERT-large-cased +2xTransformer, and we extract the fine-grained entities from the query and the retrieved articles.

4.3.2 Events. The annotated data of the ACE 2005 corpus provided by the ACE evaluation is restricted to a range of types, each with a set of subtypes. Thus, only the events of an appropriate type

¹¹ Opinions, Opinion, Letters to the Editor, The Post's View, Global Opinions, All Opinions Are Local, Local Opinions

¹² Although the organizers removed most of the duplicate articles, the process was not without faults.

¹³ <https://github.com/fmfn/BayesianOptimization>

¹⁴ This means that each query was done one by one.

¹⁵ We annotated five recommended articles per topic, about which we did not know anything. The recommended articles came from the title to title dense vector queries.

¹⁶ <https://www.elastic.co/guide/en/elasticsearch/reference/7.12/search-explain.html>

¹⁷ We explored different nDCG cuts, such as 50, 20 and 5. However, we found that, empirically, optimizing at 10 provided the best global results.

¹⁸ <https://tac.nist.gov/2020/KBP/RUFES/index.html>

¹⁹ <http://catalog.ldc.upenn.edu/LDC2006T06>

²⁰ RUFES annotation guidelines: https://tac.nist.gov/2020/KBP/RUFES/guidelines/RUFES2020AnnotationGuidelines.v1.1_draft.pdf

²¹ *PER* refers to the entity type *Person*.

Table 1: Performance of different systems for RUFES, micro-strict.

Approaches	Precision	Recall	F1
BERT-base-cased	75.4	69.4	72.3
BERT-large-cased	79.1	72.5	75.6
+ 2 × Transformer			
BERT-base-cased	75.9	69.2	72.4
BERT-large-cased	79.9	73.2	76.4

are annotated in a document. The eight event types (with 33 subtypes in parentheses) are: *Life* (*Be-Born, Marry, Divorce, Injure, Die*), *Movement* (*Transport*), *Conflict* (*Attack, Demonstrate*), etc.

Events are distinguished from their mentions in text. An event mention or a trigger is a span of text (an extent, usually a sentence) with a distinguished trigger word and zero or more arguments, which are entity mentions, timestamps, or values in the extent. Since there is nothing inherent in the task that requires the two levels of type and subtype, we will refer to the combination of event type and subtype (e.g., *Life.Die*) as the event type. If we consider the example sentence “*There was the free press in Qatar, Al Jazeera but its’ offices in Kabul and Baghdad were bombed by Americans.*”, an event extractor should detect a *Conflict.Attack* event mention, with the trigger word *bombed*.

For detecting events, we focus on the event mention detection, and we use a BERT-based model with entity markers [2, 6, 21, 22]. This method is adapted from the BERT-based model with *EntityMarkers* [2] applied for relation classification, to perform event detection.

The *EntityMarkers* model consists in augmenting the input data with a series of special tokens, e.g., if we consider a sentence $x = [x_0, x_1, \dots, x_n]$ with n tokens, we augment x with two reserved word pieces to mark the beginning and the end of each entity in the sentence. Thus, the previous sentence becomes: *There was the free press in [GPE.Country_{start}] Qatar [GPE.Country_{end}], [ORG.CommercialOrganization_{start}] Al Jazeera [ORG.CommercialOrganization_{end}] but its’ offices in [GPE.City_{start}] Kabul [GPE.City_{end}] and Baghdad were bombed by [ORG.Government.Agency_{start}] Americans [ORG.Government.Agency_{end}], where the different hierarchical entity types were detected by the previously presented model for fine-grained entity recognition.*

In Table 2, we explore again the two pre-trained and fine-tuned BERT *cased* models, the BERT-base and BERT-large, with and without the entities previously predicted. We further consider the BERT-large-cased + 2 × Transformer, and we extract the event triggers from the query and the retrieved articles.

4.3.3 Re-ranking. For each sentence of the article, the entities and the event triggers are extracted and concatenated separated by a space, forming two separate text lines. Each line of entities or event triggers is encoded with Sentence BERT and then, the final representation is the sum of all the obtained vectors $v = (v_i)_{i=1}^n$ where each element $v_{i,j} = \sum_{j=1}^n x_i, j$. We use the cosine similarity for

Table 2: Performance of different systems for ACE 2005 on the blind test data, micro-strict.

Models	Precision	Recall	F1
BERT-base-cased	71.3	72.0	71.6
BERT-large-cased	69.3	77.1	73.0
+EntityMarkers			
BERT-base-cased	79.1	72.5	75.6
BERT-large-cased	82.4	75.7	78.9

comparing the entity representations, which is defined as follows:

$$\cos(Q, R) = \frac{QR}{\|Q\| \|R\|} = \frac{\sum_{i=1}^n Q_i R_i}{\sqrt{\sum_{i=1}^n (Q_i)^2} \sqrt{\sum_{i=1}^n (R_i)^2}} \quad (5)$$

where Q is the vector representation of the Query and Retrieved is the vector representation of the retrieved article.

$$\text{score}(R) = \left(\cos(Q_{Entities}, R_{Entities}) + \cos(Q_{Events}, R_{Events}) \right) / 2 \quad (6)$$

4.4 Run 4: 300K_ENT_PH_DN

This run is a re-ranking of the Run 3 (300K_ENT_PH) (Section 4.3) in which we include the cosine distances between the article text and the description and the narrative.

$$\text{score}(R) = \left(\cos(Q_{Entities}, R_{Entities}) + \cos(Q_{Events}, R_{Events}) + \cos(Q_{BodyText}, R_{Narrative}) + \cos(Q_{BodyText}, R_{Description}) \right) / 4 \quad (7)$$

4.5 Run 5: Lambda_narr

This run consisted in starting from the outcome produced by the Lambda approach (Section 4.2) and re-ranking the recommended articles using the narrative field. The narrative field is an element provided by TREC organizers, as shown in Figure 1. It offers a summary of what background is expected.

First, we calculated the cosine similarity between the narrative field dense vector and the recommended article’s body dense vector. Then, we used a weighted harmonic mean to merge the rankings produced by the cosine similarity (R_{Narr}) and those produced by the Lambda approach (R_{Lambda}):

$$\text{Lambda_narr} = \frac{3.25 R_{Lambda}^{-1} R_{Narr}^{-1}}{(2.25 R_{Lambda}^{-1}) + R_{Narr}^{-1}} \quad (8)$$

We used the reciprocal of all the rankings R , to indicate that the lower the rankings, i.e. 1st, the better. In Equation 8 we give priority to the ranking produced by R_{Narr} over R_{Lambda} .

To produce the final ranking, we sort *Lambda_narr* scores in descending order.

4.6 Subtopics Approaches

Regarding the background of articles following the subtopics, we submitted five different approaches, that are an extension of the previously described ones.

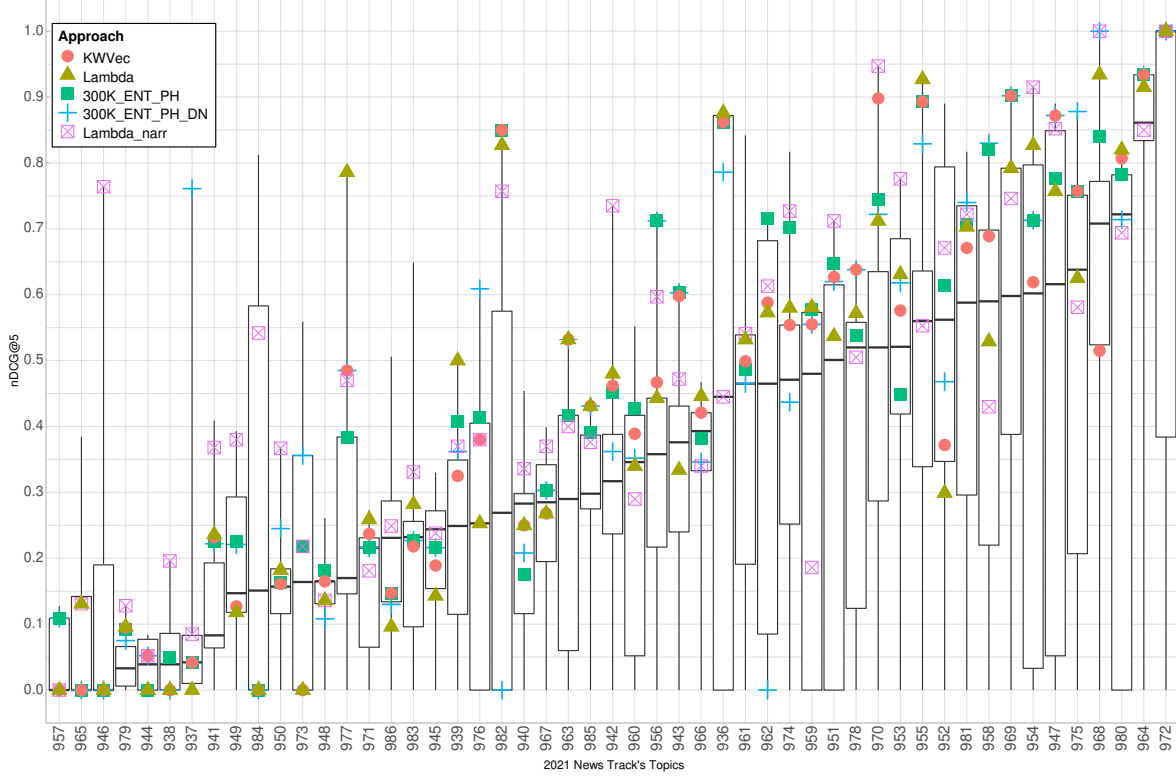


Figure 2: Boxplots of nDCG@5 score distribution for each topic based on all News Track submissions. The topics are sorted by their median nDCG@5. We present as well the nDCG@5 scores gotten by each of our approaches.

Run 1 (KWVec_sub): For this approach, we made use of the ranking produced by KWVec (Section 4.1) and re-ranked the recommended articles according to their cosine similarity with the subtopic. The re-ranking was done using the same ideas used in Section 4.5. Similarly, we applied a modified version of Equation 8:

$$KWVec_sub = \frac{3.25R_{KWVec}^{-1}R_{subtopic}^{-1}}{(2.25R_{KWVec}^{-1}) + R_{subtopic}^{-1}} \quad (9)$$

Run 2 (Lambda_sub): This run is similar to KWVec_sub. However, instead of using the outcomes produced by KWVec, we make use of the outcomes produced by Lambda (Section 4.2). We also use Equation 9 with the respective changes to use R_{Lambda} instead of R_{KWVec} .

Runs 3, 4, & 5: Run 3 is a re-ranking of the initial runs to which the cosine similarity between the text body of the query article and the text of the subtopic are added. Runs 4 and 5 have the entities and

the events removed, respectively.

$$score(R) = \left(\cos(Q_{Entities}, R_{Entities}) + \cos(Q_{Events}, R_{Events}) + \cos(Q_{BodyText}, R_{Narrative}) + \cos(Q_{BodyText}, R_{Description}) + \cos(Q_{BodyText}, R_{SubtopicText}) \right) / 5 \quad (10)$$

5 RESULTS

In Figure 2, we present, for each 2021 topic, the distribution of nDCG@5 scores calculated from all the submissions along with the scores obtained by each of our approaches. We can notice that for some topics, e.g. 957 or 979, it was very hard to predict a good background article for all the participants. In these cases, the median is not only very low, but the full distribution is quite compact and close to zero. This contrasts with other topics, like 946, 937 and 977, where despite having a low median, at least one of our approaches managed to reach values similar or equal to the maximum nDCG@5 score. Finally, we can observe that for some topics it was easy to predict background articles for most participants, such as topic 964 and 972.

In Table 3 we present a summary of Figure 2, where we indicate the number of nDCG@5 scores, produced by our runs for each topic, found within each nDCG@5 quartiles. It should be noted, that in Table 3, if the value associated with a quartile was equal to another one, e.g. $Q_0 = Q_1$, like in topic 946, the score was assigned to the quartile closest to the median one (Q_2).

Based on the results present in Table 3, we can determine that the recommendations produced by our approaches generated an nDCG@5 greater than the participants' median in at least 60% of the topics. Specifically, KWVec 66.6%, Lambda 60.7%, 300K_ENT_PH 74.5%, 300K_ENT_PH_DN 64.7% and Lambda_narr 70.5%. Moreover, all our approaches achieved the maximum score nDCG@5 score in at least 9.8% of the topics, topped by 300K_ENT_PH_DN with a 21.5%.

In Figure 3, we present the distribution of nDCG@5 scores generated by each of our explored approaches. We can notice in Figure 3, that the best system has been KWVec with an nDCG@5 median of 0.462. We can further observe that for KWVec and Lambda the distribution of the scores tends to be negatively skewed, while 300K_ENT_PH, 300K_ENT_PH_DN, and Lambda_narr are positively skewed.

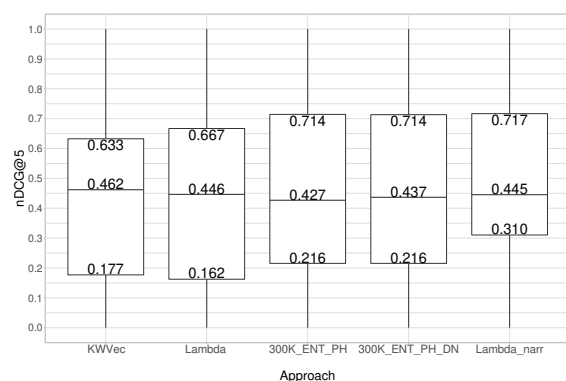


Figure 3: Boxplots representing the distribution of nDCG@5 scores obtained by each explored approach. We include the numerical values for the first, second (median), and third quartiles.

6 DISCUSSION

One aspect that we noticed from KWVec during the experimentation with the 2020 topics is that the scores obtained by the cosine similarity were, in multiple cases, diminished by the scores obtained by keywords. In other words, the final score given by ODFE to a document came mostly from the keywords, and not from the cosine similarity calculations. This is why we added weight (250) to Equation 2. However, this number was chosen experimentally based on 2020 topics.

Due to this, we decided to explore the Lambda approach, where we expected that the Bayesian optimization could automatically determine the weights (λ) that should be used to merge the scores to get the best nDCG scores. Nonetheless, the performance of Lambda

did not surpass that of KWVec, even if similar queries were used along with more specific ones.

There are multiple possible reasons why the Lambda approach did not surpass KWVec's performance. In the first place, for training, we relied on data from previous years which were produced using different methods. This means that for training we used documents that on occasions would not be retrieved by our queries as highly relevant, and therefore we introduce a bias in the weights of certain queries. Sometimes the top retrieved documents by our queries had to be removed from the training as we did not know their gold standard relevance. In spite of the fact that we manually annotated some top retrieved documents, for which we did not have a gold standard relevance score, the additional scored documents seemed to be insufficient for the training. This last point can be because of the annotation quality and variety, as it focused on one type of query, the title-title similarity, and the process was done by just one person, who could naturally be biased.

With respect to Lambda_narr, although it did not surpass KWVec performance, we can determine from Figure 3, that re-ranking the documents according to the narrative produced interesting results. We managed to set 50% of the nDCG@5 scores within a smaller and better range of values [0.310, 0.717] with respect to the other approaches. Nonetheless, most of the Lambda_narr scores were closer to 0.310 rather than to 0.717, creating a positively skewed distribution that affected its median. Despite this, Lambda_narr's median, 0.445, is similar to the one set by its parent, the Lambda approach, with an nDCG@5 of 0.446.

In regard to the re-rankings enhanced with entities and events or narratives, both runs, 300K_ENT_PH and 300K_PH_DN are rather homogeneous, with the same range of values [0.126, 0.714], and slightly similar median values. However, both Q_1 and Q_3 nDCG@5 scores surpass those of KWVec and Lambda.

It is interesting to observe that despite the fact that the model 300K_PH_DN achieved the largest number of topics with a maximum score, 11 as seen in Table 3, its median did not surpass KWVec. It is possible that the 300K_PH_DN median was severely affected by the nDCG@5 scores of topics 982 and 962, which were zero, as seen in Figure 2.

In all the cases, the results obtained by 300K_ENT_PH and 300K_PH_DN, and especially the latter, could indicate that background linking could benefit from augmenting the articles with additional extracted information, such as named entities and events.

7 CONCLUSION

In this work, we presented the participation of the *Laboratory L3i, University of La Rochelle*, at TREC 2021 News Track Background Linking. From our participation, we noticed that, despite the existence of embeddings from fine-tuned language models such as Sentence BERT [28], keywords are still one of the most powerful sources of knowledge to rank news articles. Also, we observed that extracting additional textual elements, such as named entities and events, can be useful and, in some cases, they can provide unique information that will bring out the most relevant articles. Furthermore, re-ranking news articles based on simple inputs from journalists, like a summary of what it is expected to retrieve, can improve the performance of a news background linking system. Regarding training a

Table 3: Number of topics' nDCG@5 score found in each topic's quartile (Q) calculated by TREC organizers. The value in brackets represents the percentage of topics. Q_0 is the minimum score, Q_2 is the median and Q_4 is the maximum score.

Run	$x = Q_0$	$Q_0 < x < Q_1$	$Q_1 \leq x < Q_2$	$x = Q_2$	$Q_2 < x \leq Q_3$	$Q_3 < Q_4$	$x = Q_4$
KWVec	0 (0.0)	1 (1.9)	10 (19.6)	6 (11.7)	16 (31.3)	13 (25.4)	5 (9.8)
Lambda	1 (1.9)	3 (5.8)	12 (23.5)	4 (7.8)	11 (21.5)	13 (25.4)	7 (13.7)
300K_ENT_PH	0 (0.0)	0 (0.0)	8 (15.6)	5 (9.8)	17 (33.3)	16 (31.3)	5 (9.8)
300K_ENT_PH_DN	1 (1.9)	2 (3.9)	10 (19.6)	5 (9.8)	12 (23.5)	10 (19.6)	11 (21.5)
Lambda_narr	0 (0.0)	0 (0.0)	12 (23.5)	3 (5.8)	12 (23.5)	14 (27.4)	10 (19.6)

model which optimizes weights of different queries is still difficult. Nonetheless, based on our results, it could be feasible, but more annotated data would be necessary to reduce bias.

Finally, as future work, we would like to apply the previously explored background linking approaches in less-represented languages, such as Croatian and Finnish, through the Embeddia project [26].

ACKNOWLEDGMENTS

This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (News-Eye) and 825153 (Embeddia), and by the ANNA and Termitrad projects funded by the Nouvelle-Aquitaine Region.

REFERENCES

- [1] Ahmet Aker, Emina Kurtic, Mark Hepple, Rob Gaizauskas, and Giuseppe Di Fabrizio. 2015. Comment-to-Article Linking in the Online News Domain. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czech Republic, 245–249. <https://doi.org/10.18653/v1/W15-4635>
- [2] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2895–2905. <https://doi.org/10.18653/v1/P19-1279>
- [3] Emanuela Boros and Antoine Doucet. 2021. Transformer-based Methods for Recognizing Ultra Fine-grained Entities (RUFES). *arXiv preprint arXiv:2104.06048* (2021).
- [4] Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, and Antoine Doucet. 2020. Alleviating Digitization Errors in Named Entity Recognition for Historical Documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL)*, Raquel Fernández and Tal Linzen (Eds.). Association for Computational Linguistics, Online, 431–441. <https://doi.org/10.18653/v1/2020.conll-1.35>
- [5] Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose Moreno, Nicolas Sidere, and Antoine Doucet. 2021. Atténuer les erreurs de numérisation dans la reconnaissance d'entités nommées pour les documents historiques. In *Conférence en Recherche d'Informations et Applications-CORIA 2021, French Information Retrieval Conference*. CORIA, Online.
- [6] Emanuela Boros, Jose G. Moreno, and Antoine Doucet. 2021. Event Detection with Entity Markers. In *Advances in Information Retrieval*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer International Publishing, Cham, 233–240.
- [7] Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, José Moreno, Nicolas Sidere, and Antoine Doucet. 2020. Robust named entity recognition and linking on historical multilingual documents. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névelo (Eds.), Vol. 2696. CEUR-WS, Thessaloniki, Greece, 1–17.
- [8] Luis Adrián Cabrera-Diego, Stéphane Huet, Bassam Jabaian, Alejandro Molina, Juan-Manuel Torres-Moreno, Marc El-Bèze, and Barthélémy Durette. 2014. Algorithmes de classification et d'optimisation : participation du LIA/ADOC à DEFT'14. In *Actes du dixième Défi Fouille de Textes*. Association pour le Traitement Automatique des Langues, Marseille, France, 53–60.
- [9] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célio Nunes, and Adam Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences* 509 (2020), 257–289. <https://doi.org/10.1016/j.ins.2019.09.013>
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, USA, 4171–4186. <https://doi.org/10.18653/v1/n19-1423>
- [11] Elisa Shearer. 2021. More than eight-in-ten Americans get news from digital devices. *Pew Research Center* (Dec. 2021). <https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/>
- [12] Marwa Essam and Tamer Elsayed. 2020. Why is That a Background Article: A Qualitative Analysis of Relevance for News Background Linking. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (Virtual Event, Ireland) (CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 2009–2012. <https://doi.org/10.1145/3340531.3412120>
- [13] Galen Stocking and Maya Khuzam. 2021. Digital News Fact Sheet. *Pew Research Center* (June 2021). <https://www.pewresearch.org/journalism/fact-sheet/digital-news/>
- [14] Weiwei Guo, Hao Li, Heng Ji, and Mona Diab. 2013. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 239–249. <https://aclanthology.org/P13-1024>
- [15] Heng Ji, Avirup Sil, Hoa Trang Dang, Ian Soboroff, Joel Nothman, and Sydney Informatics Hub. 2019. Overview of TAC-KBP2019 Fine-grained Entity Extraction. In *2019 Text Analysis Conference Proceedings*. NIST, Gaithersburg, Maryland, USA.
- [16] Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. 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, Brussels, Belgium, 133–142. <https://doi.org/10.18653/v1/K18-2013>
- [17] Pavel Khloponin and Leila Kosseim. 2020. The CLaC System at the TREC 2020 News Track. In *Proceedings of the 29th Text REtrieval Conference (TREC 2020)*, Ellen M. Voorhees and Angela Ellis (Eds.), Vol. SP 1266. NIST, Online. <https://trec.nist.gov/pubs/trec29/papers/CLaC.N.pdf>
- [18] Pavel Khloponin and Leila Kosseim. 2021. Using Document Embeddings for Background Linking of News Articles. In *Natural Language Processing and Information Systems*, Elisabeth Métais, Farid Meziane, Helmut Horacek, and Epaminondas Kapetanios (Eds.). Springer International Publishing, Cham, 317–329.
- [19] Boshko Koloski, Elaine Zosa, Timen Stepišnik-Perdih, Blaž Škrli, Tarmo Paju, and Senja Pollak. 2021. Interesting cross-border news discovery using cross-lingual article linking and document similarity. In *Proceedings of the EACL Hackshop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics, Online, 116–120. <https://aclanthology.org/2021.hackashop-1.16>
- [20] Kuang Lu and Hui Fang. 2019. Leveraging Entities in Background Document Retrieval for News Articles. In *Proceedings of the 28th Text REtrieval Conference (TREC 2019)*, Ellen M. Voorhees and Angela Ellis (Eds.), Vol. SP 1250. NIST, Online. <https://trec.nist.gov/pubs/trec28/trec2019.html>
- [21] Jose G Moreno, Emanuela Boros, and Antoine Doucet. 2020. TLR at the NTCIR-15 FinNum-2 Task: Improving Text Classifiers for Numeral Attachment in Financial Social Data. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*. National Institute of Informatics, Tokyo, Japan, 8–11.
- [22] José G. Moreno, Antoine Doucet, and Brigitte Grau. 2021. Relation Classification via Relation Validation. In *Proceedings of the 6th Workshop on Semantic Deep*

TREC, 15-19 November 2021, Online

Cabrera-Diego, et al.

- Learning (SemDeep-6)*. Association for Computational Linguistics, Online, 20-27. <https://aclanthology.org/2021.semdeep-1.4>
- [23] Jonas Močkus, Vytautas Tiešis, and Antanas Žilinskas. 1978. The application of Bayesian methods for seeking the extremum. In *Towards Global Optimisation*, George Philip Szegő and Laurence Charles Ward Dixon (Eds.), Vol. 2. North-Holland, Amsterdam, The Netherlands, 117-128.
- [24] Shahrzad Naseri, John Foley, and James Allan. 2018. UMass at TREC 2018: CAR, Common Core and News Tracks. In *Proceedings of the 27th Text REtrieval Conference (TREC 2018)*, Ellen M. Voorhees and Angela Ellis (Eds.), Vol. SP 500-331. NIST, Gaithersburg, Maryland, USA. <https://trec.nist.gov/pubs/trec27/papers/Overview-News.pdf>
- [25] Joel Nothman, Matthew Honnibal, Ben Hachey, and James R. Curran. 2012. Event Linking: Grounding Event Reference in a News Archive. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Jeju Island, Korea, 228-232. <https://aclanthology.org/P12-2045>
- [26] Senja Pollak, Marko Robnik-Šikonja, Matthew Purver, Michele Boggia, Ravi Shekhar, Marko Pranjic, 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 Škrj, Martin Žnidaršič, Andraž Pelicon, Boshko Koloski, Vid Podpečan, Janez Kranjc, Shane Sheehan, Emanuela Boros, Jose G. 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, Online, 99-109. <https://www.aclweb.org/anthology/2021.hackashop-1.14>
- [27] Marko Pranjic, Vid Podpečan, Marko Robnik-Šikonja, and Senja Pollak. 2020. Evaluation of related news recommendations using document similarity methods. In *Proceedings of the Conference on Language Technologies and Digital Humanities (JDT2020)*, Darja Fišer and Tomaž Erjavec (Eds.), Inštitut za novejšo zgodovino, Ljubljana, Slovenia, 81-86. <https://doi.org/10.5281/zenodo.4059710>
- [28] 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)*. Association for Computational Linguistics, Hong Kong, China, 3982-3992. <https://doi.org/10.18653/v1/D19-1410>
- [29] Shudong Huang, Ian Soboroff, and Donna Harman. 2018. TREC 2018 News Track. In *Proceedings of the Second International Workshop on Recent Trends in News Information Retrieval (NewsIR'18)*, Dyaa Albakour, David Corney, Julio Gonzalo, Miguel Martinez, Barbara Poblete, and Andreas Valochas (Eds.), Vol. 2079. CEUR Workshop Proceedings, Grenoble, France. <http://ceur-ws.org/Vol-2079/paper12.pdf>
- [30] Ian Soboroff, Shudong Huang, and Donna Harman. 2018. TREC 2018 News Track Overview. In *Proceedings of the 27th Text REtrieval Conference (TREC 2018)*, Ellen M. Voorhees and Angela Ellis (Eds.), Vol. SP 500-331. NIST, Gaithersburg, Maryland, USA. <https://trec.nist.gov/pubs/trec27/papers/Overview-News.pdf>
- [31] Ian Soboroff, Shudong Huang, and Donna Harman. 2020. TREC 2020 News Track Overview. In *Proceedings of the 29th Text REtrieval Conference (TREC 2020)*, Ellen M. Voorhees and Angela Ellis (Eds.), Vol. SP 1266. NIST, Online. <https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.N.pdf>
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., Long Beach, CA, USA, 5998-6008.
- [33] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A Large-scale Dataset for News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 3597-3606. <https://doi.org/10.18653/v1/2020.acl-main.331>