News Linker
Project Proposal for NLP Course, Winter 2023

Team Member 1: Panpan Liu
Warsaw University of Technology
01183030@pw.edu.pl

Team Member 2: Trifebi Shina Sabrila
Warsaw University of Technology
01185877@pw.edu.pl

Team Member 3: Illia Tesliuk
Warsaw University of Technology
01138770@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

Nowadays, the ability to seamlessly connect related pieces of information across different mediums is essential for coherent news dissemination. The "News Linker" project introduces an innovative system designed to automatically links entities across a variety of data formats, such as text articles, images, audio streams, and video interviews, focused on the same news events. By leveraging state-of-the-art Natural Language Processing (NLP) techniques and entity linking frameworks, the system identifies and unifies related entities and topics across different media, enabling a seamless narrative flow. This not only enrich the news experience for both content creators and consumers but also streamlines editorial workflows. Focusing on STA's Slovenian news data, our project aims to deliver a proof-of-concept system that enhances the coherence of news reporting.

1 Introduction

The advent of digital media has transformed the way news is reported. The deluge of news is no longer limited to traditional text-based articles. Press agencies now deliver news in a multimedia format, ranging from text articles to images, audio and video interviews. The growth of these diverse media formats often leads to a fragmentation that prevents journalists from efficiently tracking news developments and crafting comprehensive narratives. This not only impacts internal editorial operations but also disrupts the news experience for audiences who want a diverse understanding of news events.

In response to these challenges, our 'News Linker' project seeks to bridge the gap between these various forms of news data. Rather than focusing solely on linking news to external sources like social media platforms (Mogadala et al., 2017) or research journals (Wang and Yu, 2021) as has been done in previous studies, our project takes a unique approach. We prioritize on creating an automatic linking system that brings different types of news data around the same news topics or events within this diverse landscape of news data.

To achieve this, we utilize the power of Named Entity Recognition (NER) techniques, extracting valuable metadata from STA's Slovenian language news data. This metadata serves as the cornerstone of our entity linking process which then will be used on the entity linking process by the use of some pretrained models. This system is designed to reduce the time journalists spend on cross-referencing and validating information, leading to a more coherent and engaging story experience for the end-user.

The primary goal of the 'News Linker' project is to develop a proof-of-concept system that can accurately and efficiently link related content within STA's Slovenian news data. By focusing on this dataset, we aim to showcase the potential of automatic news data linking.

The research question we want to address is whether the implementation of Named Entity Recognition (NER) and entity linking techniques from Natural Language Processing (NLP) advancements enable the creation of an effective system to automatically connect related multimedia news content within STA's Slovenian news data.

2 Significance of the project

The News Linker project addresses a specific and complex scientific problem: integrating text data, images, audio, video, and metadata from press agencies to create a coherent and unified understanding of news events. This problem is critical as it directly impacts the accessibility and comprehensibility of news information. Furthermore, it has real-world applications in improving news aggregation, recommendation systems, and crisis management. More specifically, the project contributes to a deeper understanding of how multimedia content can be linked to enhance event comprehension, thereby advancing the state of the art in the field.Besides, it introduces novel methods for integrating metadata and handling multilingual data, which can potentially inspire methodological advancements in cross-media linking.

3 Literature Review

One research field closely related to our project is Multilingual Text Matching. In the traditional field of machine learning, common text matching algorithms involve extracting features from text using methods such as TF-IDF (Martineau and Finin, 2009), Word2Vec (Mikolov et al., 2013), edit distance, and other linguistic characteristics. Subsequently, these features are fed into a machine learning model (e.g., logistic regression) or a statistical measure (e.g., cosine similarity) to estimate how similar the two texts are. However, this approach is relatively coarse and struggles to capture the actual semantic information in the text.

Large corpora and deep neural networks have provided language models with the ability to understand the deep-seated semantic information within sentences, thus improving the text matching algorithm's capability to calculate the semantic relevance between two pieces of text. Three popular pre-trained language models are widely used today and have become the technical backbone for many subsequent NLP tasks. The Transformer model (Vaswani et al., 2017) introduced by Google in 2017 has outperformed models like RNN and LSTM in many NLP field; Bidirectional Encoder Representations from Transformers (BERT) architecture (Devlin et al., 2018) proposed by Google in 2018 adapt to downstream tasks enable us to add a task-specific output layer to the model and fine-tune it; an improved training method for the BERT model called RoBERTa (Liu et al., 2019).

Due to the significant achievements of pre-trained language models in the English domain, researchers have started to explore the extension of this training approach to multiple languages. The goal is to use a single language model to deal with texts in various languages. Pires and his colleagues introduced Multilingual BERT (Pires et al., 2019) and conducted research on its effectiveness. The main idea of Multilingual BERT is to adapt the training strategy from BERT, using a single model's weights to handle all target languages. Because the Multilingual BERT model did not leverage information from parallel corpus sentences that are translations of each other, Lample and his colleagues subsequently introduced the XLM model (Lample and Conneau, 2019). The training of mBERT and XLM models relies on Wikipedia corpora. However, Conneau and his colleagues found that this approach was not very supportive of low-resource languages. Consequently, they balanced the high-resource and low-resource languages in the training data and, following the XLM model's methodology, introduced the XLM-R model (Conneau et al., 2019).

4 Concept and work plan

This section provides an overview of the project analysis and the associated timeline. It highlights the key milestones and objectives.

4.1 Project activities and timeline

The project is divided into 3 main parts, as presented in the Table 1.

Date	Stage Name
8.11.23	Project proposal
22.11.23	Proof of concept
13.12.23	Final Project

Table 1: Project activity and timeline.

4.2 Specific research goals

The following research goals are established for the project:

- gaining comprehensive knowledge of the latest advancements in the NLP domain, with a specific focus on Named Entity Recognition, Entity Linking and Topic Discovery

- testing different available NLP approaches tasks to solve the news linking problem

- working with low-resource language, adaptive high-resource-based model for these needs

- combining different NER, EL and Seed-guided topic modeling into one working solution

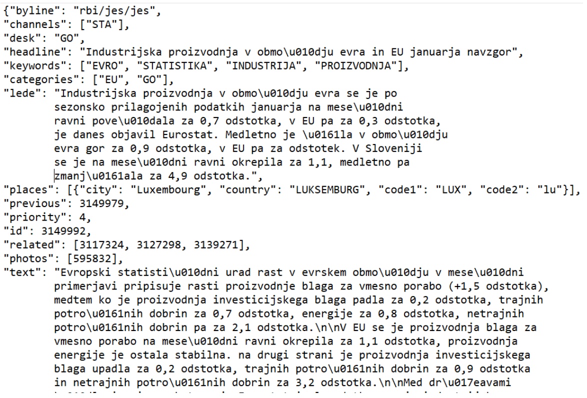
5 Approach

5.1 Datasets

News Linker project will be working on text data written in Slovenian. Our primary data source is the API of the Slovenian Press Agency (STA) - the leading provider of media content in Slovenia.

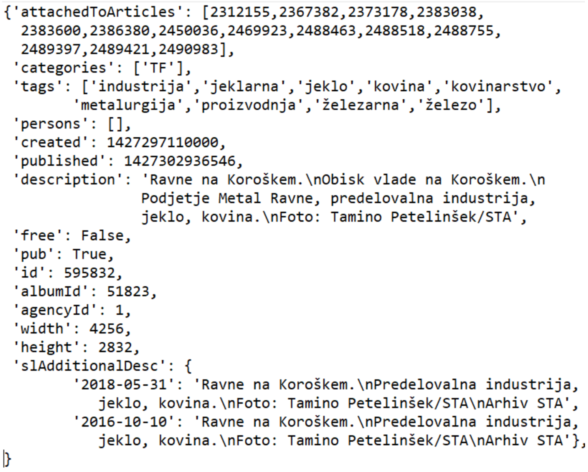
STA API enables obtaining an ID list of the articles published on a specific date in Slovenian or English. Each news article can be retrieved by sending a request with its ID number. A typical response contains the article's full text, headline, leading paragraph (lede), category, and a list of keywords. Additionally, it may also have authors' names, creation and publication dates, news priority, a list of places, and a list of related news. Some news also contains IDs of the attached photos or video albums. We are interested in their text descriptions. These medias can be retrieved by sending a request to the API with a particular image or video album ID. Each image or video record also contains a list of news IDs to which they are attached to. Therefore, we can retrieve an information about the news-media connection from either side.

We are planning to fill our text corpus with articles coming from a specific period (e.g. 3 months). However, some news may be traffic information, daily bulletins, or daily digests. The latter contains the same information as the individual articles, but in a condensed form. These categories are useless for our project and have to be filtered out at the pre-processing stage. Since the topic of our project is the integration of different types of data, we are also going to download images and video albums from a specific period and fill our corpus with their descriptions.



```
{"byline": "rbi/jes/jes",
"channels": ["STA"],
"desk": "GO",
"headline": "Industrijska proizvodnja v obmo\u010dju evra in EU januarja navzgor",
"keywords": ["EVRO", "STATISTIKA", "INDUSTRIJA", "PROIZVODNJA"],
"categories": ["EU", "GO"],
"lede": "Industrijska proizvodnja v obmo\u010dju evra se je po
        sezonsko prilagojenih podatkih januarja na mese\u010dni
        ravni pove\u010dala za 0,7 odstotka, v EU pa za 0,3 odstotka,
        je danes objavil Eurostat. Medletno je \u0161la v obmo\u010dju
        evra gor za 0,9 odstotka, v EU pa za odstotek. V Sloveniji
        se je na mese\u010dni ravni okrepila za 1,1, medletno pa
        zmanj\u0161ala za 4,9 odstotka.",
"places": [{"city": "Luxembourg", "country": "LUKSEMBURG", "code1": "LUX", "code2": "lu"}],
"previous": 3149979,
"priority": 4,
"id": 3149992,
"related": [3117324, 3127298, 3139271],
"photos": [595832],
"text": "Evropski statisti\u010dni urad rast v evrskem obmo\u010dju v mese\u010dni
        primerjavi pripisuje rasti proizvodnje blaga za vmesno porabo (+1,5 odstotka),
        medtem ko je proizvodnja investicijskega blaga padla za 0,2 odstotka, trajnih
        potro\u0161nih dobrin za 0,7 odstotka, energije za 0,8 odstotka, netrajnih
        potro\u0161nih dobrin pa za 2,1 odstotka.\n\nV EU se je proizvodnja blaga za
        vmesno porabo na mese\u010dni ravni okrepila za 1,1 odstotka, proizvodnja
        energije je ostala stabilna. na drugi strani je proizvodnja investicijskega
        blaga upadla za 0,2 odstotka, trajnih potro\u0161nih dobrin za 0,9 odstotka
        in netrajnih potro\u0161nih dobrin za 3,2 odstotka.\n\nMed dr\u017eavami
```

Figure 1: An example of an STA news article



```
{'attachedToArticles': [2312155,2367382,2373178,2383038,
  2383600,2386380,2450036,2469923,2488463,2488518,2488755,
  2489397,2489421,2490983],
 'categories': ['TF'],
 'tags': ['industrija','jeklarna','jeklo','kovina','kovinarstvo',
        'metalurgija','proizvodnja','železarna','železo'],
 'persons': [],
 'created': 1427297110000,
 'published': 1427302936546,
 'description': 'Ravne na Koroškem.\nObisk vlade na Koroškem.\n
                Podjetje Metal Ravne, predelovalna industrija,
                jeklo, kovina.\nFoto: Tamino Petelinšek/STA',
 'free': False,
 'pub': True,
 'id': 595832,
 'albumId': 51823,
 'agencyId': 1,
 'width': 4256,
 'height': 2832,
 'slAdditionalDesc': {
        '2018-05-31': 'Ravne na Koroškem.\nPredelovalna industrija,
           jeklo, kovina.\nFoto: Tamino Petelinšek/STA\nArhiv STA',
        '2016-10-10': 'Ravne na Koroškem.\nPredelovalna industrija,
           jeklo, kovina.\nFoto: Tamino Petelinšek/STA\nArhiv STA'},
}
```

Figure 2: An example of an STA photo data

It's worth mentioning that all three types of data, namely, articles, images and videos are represented in a text form. However, media descriptions are usually much shorter than the article sentences and are mostly composed of named entities. Typically, they would include only 4-5 words of topic-related information and a name of the image's author or press agency. The latter information should be removed as it doesn't describe image's content.

If a user would like to add a piece of media to the corpus, an additional description annotation has to be done. The annotation can be done either manually or with the help of pre-trained models (e.g. image captioning network). Development of annotation models is out of the scope of our project. We assume that our framework receives a text corpus and a set of seeds as an input.

Apart from the data collected from an STA API, we would also like to have a labeled dataset that can be used for fine-tuning pre-trained Large Language Models. However, most of modern NLP models are trained and evaluated on high-resource languages, such as English. In turn, low-resource languages, such as Slovenian, typically have only a limited number of text corpora with labels prepared for Named Entity Recognition or Entity Linking available. For some tasks there can be no such corpora at all.

We found a Slovenian ”SUK 1.0” corpus of 2913 texts with manual annotatation prepared for different NLP tasks, including named entity recognition. Namely, ssj500k-syn (200 320 words) and SentiCoref (340 401 words) parts contain Slovene-named entities and can be used for fine-tuning English or Slovene Large Language Models.

5.2 Methods

The goal of the News Linker project is to produce a method that given some event returns from a text corpus a list of different kinds of data describing this event. Namely, a text corpus produced from STA API contains three distinct types of data - news articles, images and video volumes. The latter two contain string descriptions, therefore the whole project is done solely in a text domain.

News Linker project can be viewed as an intersection of such Natural Language Processing tasks as Named Entity Recognition (NER), Entity Linking (EL), Seed-Guided Topic Discovery or Clustering. Since there is no a single unique method for a news linking tasks, we are planning to implement and test several approaches.

Namely, the first one relies on executing an end-to-end Entity Linking or a combination of Named Entity Recognition and disambiguation-only EL on a text corpus and retrieving named entities from each document. Next, the similarity between the input term and the retrieved values has to be calculated and the IDs of the documents containing the high-score entities are returned. This approach implies using some of the state-of-the-art NER and EL models. ACE+document-context (Wang et al., 2020) and LUKE (Yamada et al., 2020) achieve remarkable 94.6% and 94.3% F1-scores on English *CoNLL 2003* task, correspondingly. However, due to language difference an additional research has to be done on the possibility of fine-tuning the above-mentioned SOTA methods on the named-entity-labeled parts of *SUK 1.0* corpus.

NER models can be used together with SOTA disambiguation-only Entity Linking models such as DeepType (Raiman and Raiman, 2018) which achieve almost 95% micro-precision score on English *AIDA CoNLL-YAGO Dataset*.

The next section describes an iterative seed-guided topic discovery framework called *SeedTopicMine* (Zhang et al., 2023). Its diagram is presented in Figure 3.

5.3 Seed-Guided Topic Discovery

Seed-guided topic discovery is a technique used in topic modeling to guide the process of identifying and extracting specific topics themes from a collection of text data with an assistance of seed terms, that serve as initial hints and direct the topic discovery process. The goal is to ensure that the generated topics are relevant to the goals of the user. In case of the News Linker project, basic event information can be used as the seeds, namely, type of event, its name, place or participants.

5.3.1 Types of Context Information

SeedTopicMine method joins three different types of context information and gradually fuses their context signals via an ensemble ranking process. This approach allows the contexts to complement each other and overcome their limitations.

The network's goal is to produce a set of terms related to every input seed. The term set starts with the seed itself and inthe consecutive iterations gradually expands with the terms close to the seed's semantic category in an embedding space.

The initial context involves Seed-Guided Text Embeddings. The concept behind learning text embeddings relies on the idea that terms with similar meanings tend to appear in similar contexts. *SeedTopicMine* integrates three distinct context types, i.e., a term's skip-gram, the documents it appears in, and the category it belongs to into a single objective. The goal is to maximize the probability of encountering a term's skip-gram considering its document and category contexts. Next, the cosine similarity between the embeddings of the retrieved term and its seed is calculated as an initial measure of their semantic closeness.

The second context utilizes Pre-trained Language Model Representations. Models such as BERT acquire extensive general knowledge from vast corpora like Wikipedia, which can complement the information within the input corpus. Since our project is focused on Slovene text data, original BERT models can't be applied to our text corpora. Therefore, we can use a pre-trained monolingual Slovene BERT-like model *SloBERTa* or a trilingual *CroSloEngual BERT* that is trained on Croatian, Slovenian, and English corpora and can be used for cross-lingual knowledge transfer. Spacy also library provides a set of pipelines of different sizes trained on news in Slovenian.

For each term within the input corpus *SeedTopicMine* feeds the sentences containing instances of the term into a pre-trained language model, aggregates the results over instances, and obtains a vector of a model-specific dimension per

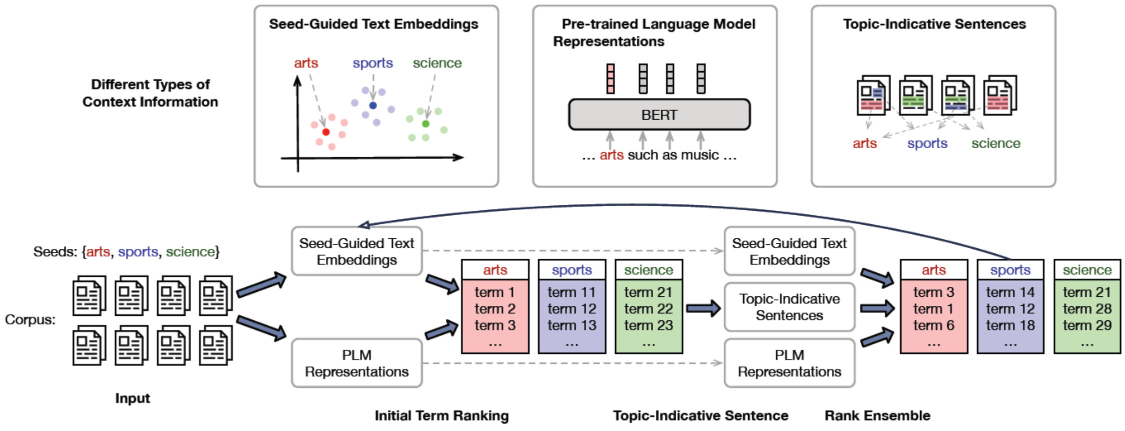

Different Types of Context Information
Seed-Guided Text Embeddings
arts
sports
science
Pre-trained Language Model Representations
BERT
... arts such as music ...
Topic-Indicative Sentences
arts
sports
science
Seeds: {arts, sports, science}
Corpus:
Input
Seed-Guided Text Embeddings
PLM Representations
arts
term 1
term 2
term 3
...
sports
term 11
term 12
term 13
...
science
term 21
term 22
term 23
...
Seed-Guided Text Embeddings
Topic-Indicative Sentences
PLM Representations
arts
term 3
term 1
term 6
...
sports
term 14
term 12
term 18
...
science
term 21
term 28
term 29
...
Initial Term Ranking
Topic-Indicative Sentence
Rank Ensemble

Figure 3: Diagram of *SeedTopicMine* framework (Zhang et al., 2023)

term. Semantic proximity between a given term and a seed is assessed based on the cosine similarity between their representations derived from the pre-trained language model.

Finally, the framework checks whether the utilized context information is topic-indicative or not. For each seed, it assumes an existence of a set of topic-indicative sentences. Assessment of the semantic proximity between a term and a sentence is based term's popularity (such term has to appear frequently in a topic-indicative sentence) and distinctiveness (the term should be much more relevant to its topic-indicative sentence compared to sentences of other topics). This relevance can be quantified using the BM25 function.

5.3.2 Initial Term Ranking

The framework starts with a single seed for each semantic category and doesn't have any topic-indicative sentences yet. It uses seed-guided text embeddings and PLM-based representations to find terms that are relevant to each category and adds the terms with the largest semantic proximity scores to the topic-indicate set of terms.

5.3.3 Topic-Indicative Sentence Retrieval

Based on the set of updated topic-indicative terms, the framework captures a set of topic-indicative sentences from the text corpus. First it retrieves sentences with all topic-indicative terms coming from only one category - so-called ”anchor” sentences. The retrieval starts with the sentences containing the largest number of terms. Next, the framework searches for the neighbors of the ”anchors”. They are added to a set of topic-indicative sentences only if their terms are coming from the same category and there's no terms of other categories.

5.3.4 Ensemble of Multiple Types of Contexts

After obtaining sets of topic-indicative sentences of each category the framework is able to calculate semantic proximity between the terms and these sentences. Next, for each term the semantic proximity with different categories is calculated. It jointly considers similarity scores for each of three types of contexts. The framework ranks the terms based on combined, PLM- and embedding-based scores. Finally, it performs rank ensemble by calculating the mean reciprocal rank (MRR) and adds the terms with MRR greater than a specified threshold to a set of topic-indicative terms. The updated set is fed to the next iteration.

Since we are interested in the documents containing information about a specific seed, *SeedTopicMine* can be slightly adjusted to save not only topic-indicative terms, but also the IDs of the corresponding documents in the text corpus.

5.3.5 Evaluation metrics

Given discovered terms under each seed we are planning evaluate the results based on *topic coherence* and *term accuracy* criteria. They are computed with the help of the following metrics:

- **NPMI** (Lau et al., 2014)serves as a commonly utilized metric in topic modeling for assessing topic coherence within each subject. It is determined by calculating the average normalized pointwise mutual information across each pair of terms

- **P@k**, also known as MACC (Mean Average Class Cohesion) (Meng et al., 2020), functions as a metric assessing term accuracy. It measures the proportion of retrieved terms that actually belong to a particular semantic category. This metric is contingent upon human judgment and requires annotations to be manually conducted by annotators. The reported **P@k** score represents the average **P@k** across all annotators

- **NDCG@k** stands as another metric for term accuracy, assigning greater importance to terms with higher ranks through the application of a logarithmic discount. Like the previous metric, human-made annotations are also necessary for this evaluation.

6 Implementation Plan

We will use STA API to collect Slovenian data over a significant period of time, e.g. 1 year. The preliminary exploratory data analysis discovered some English words in the subset collected over 3 weeks, therefore particular attention will be paid to filter out all the data that could harm the performance of our framework.

The collected data will be used it to form a text corpus and a validation set. The validation set will consist of text representations of news articles, photos, and videos as the input data. Each photo and video record contains IDs of the news articles that include it. These IDs will be added as the gold standard validation targets. We will expect these IDs to be in the outputs produced by our framework after feeding it the corresponding photo and video descriptions.

The performance of the framework will be assessed by the percentage of the gold standard targets present in the model outputs. In case the framework produces some additional IDs that are not present in the target set, we will perform a manual assessment of the answer to tell if the corresponding recourse indeed can be linked to the input photo or video description. The unrelated answers will reduce the scoring of the framework.

In turn, news articles contain links to photos and videos, so their IDs will also be used as gold standard targets. However, there are no links to other news articles. To solve this problem, we are going to use lists of keywords to find articles that can be used as proper targets. Additionally, we are planning to use the pretrained Named Entity Recognition model to extract named entities from the potential target candidates. Articles that share the same named entities will be considered linked and will be added to the target list. We are planning to test the framework with and without additional targets.

7 Conclusion

In conclusion, by utilizing a combination of Named Entity Recognition, Entity Linking and Seed-Guided Topic Discovery, this project aims to create a system that identifies and unifies related entities and topics of news across different media, making it easier for journalist and audience to track and understand the information from multiple sources to get the full picture.

References

[Conneau et al.2019] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

[Devlin et al.2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Lample and Conneau2019] Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

[Lau et al.2014] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In Shuly Wintner, Sharon Goldwater, and Stefan Riezler, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden, April. Association for Computational Linguistics.

[Liu et al.2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[Martineau and Finin2009] Justin Martineau and Tim Finin. 2009. Delta tfidf: An improved feature space for sentiment analysis. In *proceedings of the International AAAI Conference on Web and Social Media*, volume 3, pages 258–261.

[Meng et al.2020] Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and

Jiawei Han. 2020. Discriminative topic mining via category-name guided text embedding. In *Proceedings of The Web Conference 2020*, pages 2121–2132.

[Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Mogadala et al.2017] Aditya Mogadala, Dominik Jung, and Achim Rettinger. 2017. Linking tweets with monolingual and cross-lingual news using transformed word embeddings. *arXiv preprint arXiv:1710.09137*.

[Pires et al.2019] Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

[Raiman and Raiman2018] Jonathan Raiman and Olivier Raiman. 2018. Deeptype: multilingual entity linking by neural type system evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

[Vaswani et al.2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

[Wang and Yu2021] Jun Wang and Bei Yu. 2021. Linking health news to research literature. *arXiv preprint arXiv:2107.06472*.

[Wang et al.2020] Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2020. Automated concatenation of embeddings for structured prediction. *arXiv preprint arXiv:2010.05006*.

[Yamada et al.2020] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

[Zhang et al.2023] Yu Zhang, Yunyi Zhang, Martin Michalski, Yucheng Jiang, Yu Meng, and Jiawei Han. 2023. Effective seed-guided topic discovery by integrating multiple types of contexts. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 429–437.