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 link 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 enriches 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 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 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 will then be used on the entity-linking process using 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 fields; 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 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, an adaptive high-resource-based model for these needs

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

5 Approach

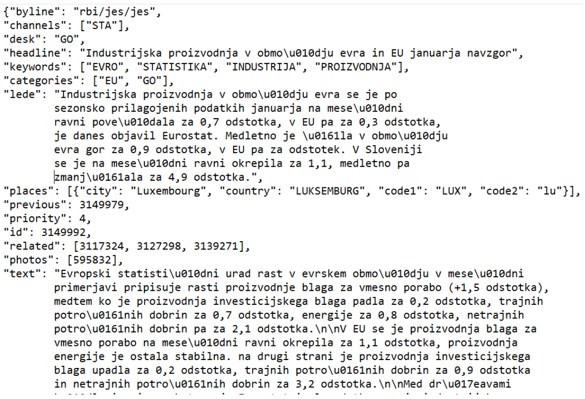
5.1 Datasets

The News Linker project focuses on analyzing text data in the Slovenian language. The primary data source for this project is the API of the Slovenian Press Agency (STA), which stands as the foremost provider of media content in Slovenia.

Utilizing the STA API, the project allows the retrieval of an ID list of articles published on a specific date in either Slovenian or English. By sending a request with the article's ID number, each news article can be obtained, accompanied by its full text, headline, leading paragraph (lede), category, and a list of keywords. Additionally, the response may include authors' names, creation and publication dates, news priority, a list of places, and related news. Some news articles may also feature IDs of attached photos or video albums, and the project specifically focuses on extracting text descriptions from these media.

To access information about attached images or video albums, a request can be sent to the API with a specific image or video album ID. Each image or video record contains a list of news IDs to which they are attached. Consequently, the News Linker project is designed to retrieve and analyze information regarding the connection between news articles and media content from either perspective.

Our strategy involves populating our text corpus with articles dated within a specified period, namely 12 months. However, it's worth noting that certain articles fall into categories such as traffic information, daily bulletins, or condensed daily digests. As these categories do not align with the objectives of our project, they need to be filtered out during the pre-processing stage.



```
{"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
```

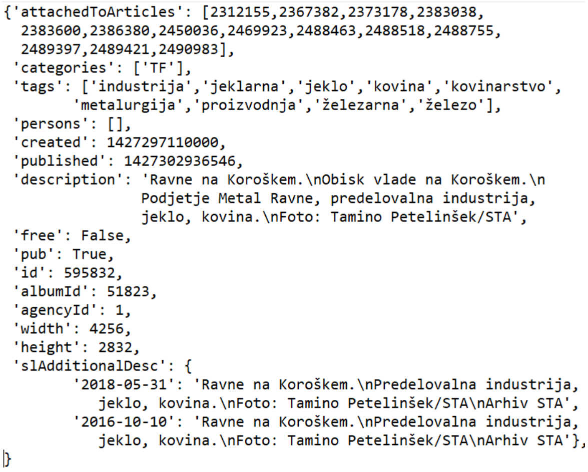
Figure 1: An example of an STA news article

Given that our project revolves around the integration of diverse data types, we are not limited to textual content alone. We plan to extend our corpus by downloading images and video albums from the same designated period and incorporating their respective descriptions. This approach ensures a comprehensive inclusion of various data formats, aligning with the overarching theme of integrating different types of data in our project.

It is important to note that the three data types—articles, images, and videos—are all presented in a textual format. However, media descriptions differ significantly from article sentences; they are typically much shorter and predominantly consist of named entities. These descriptions usually encompass only 4-5 words conveying topic-related information along with the name of the image's author or press agency. Notably, the latter information, pertaining to the author or agency, is considered irrelevant as it does not contribute to describing the image's content and should be excluded.

If a user wishes to augment the corpus with a piece of media, an additional description annotation is required. This annotation process can be carried out manually or facilitated by pre-trained models, such as an image captioning network. Development of annotation models is out of the scope of our project. An additional annotated piece of media can indeed be employed in the task of topic discovery, although it is insufficient for the task of document linking. To accomplish document linking, a user must identify articles within the primary STA dataset that can be associated with the given piece of media and subsequently attach the corresponding IDs to the media record. A pair of media annotation and a list of related articles can be used at the evaluation stage. As these operations require proficiency in Slovenian, we have decided to focus exclusively on data extracted from the STA API, avoiding the inclusion of manually added articles, photos, or videos in our approach.

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 modern NLP models are trained and evaluated on high-resource languages, such as English, while low-resource lan-



```
{'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

guages, 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 annotation 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 single unique method for news linking tasks, we plan 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 differences, 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*.

Alternatively, we would like to test 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 employed in topic modeling to facilitate the identification and extraction of specific thematic topics from a collection of textual data, utilizing seed terms as guiding indicators. These seed terms serve as initial hints, directing the topic discovery process. The goal is to ensure that the generated topics align with the user's objectives. In the context of the News Linker project, fundamental event information, such as the event type, name, location, or participants, can be employed as seeds.

We are utilizing the *SeedTopicMine* framework to extract not only terms related to the input seeds but also the documents that contain these terms.

5.3.1 Types of Context Information

SeedTopicMine method integrates three distinct types of contextual information and systematically combines their contextual signals through an ensemble ranking process. This approach enables the contexts to mutually complement each other and overcome their inherent limitations.

The network is designed to generate a set of terms related to each input seed. The term set begins with the seed itself, and in subsequent iterations, it progressively expands to include terms closely associated with the seed's semantic category in an embedding space.

The initial context involves *Seed-Guided Text Embeddings*. The concept underlying the learning

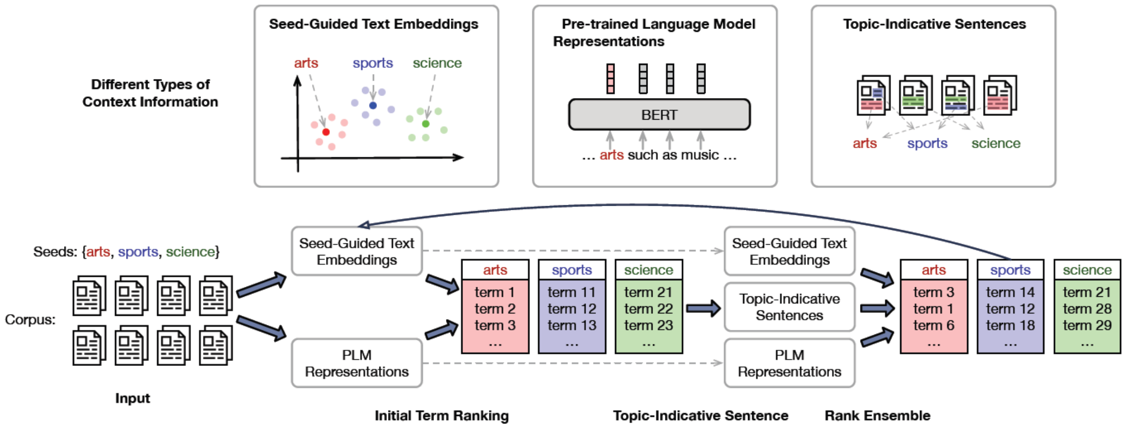

Different Types of Context Information
Seed-Guided Text Embeddings
arts
sports
science
Pre-trained Language Model Representations
BERT
... arts such as music ...
Topic-Indicative Sentences
arts
sports
science
Seeds: {arts, sports, science}
Corpus:
Input
Seed-Guided Text Embeddings
PLM Representations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

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

of text embeddings relies on the notion that terms with similar meanings tend to appear in analogous contexts. *SeedTopicMine* integrates three distinct types of context: a term's skip-gram, the documents it appears in, and the category to which it belongs, into a unified objective. The aim is to maximize the probability of encountering a term's skip-gram while taking into account its document and category contexts.

The skip-gram objective is defined as:

$$J_{skip} = log \prod_{d \in D} \prod_{w_i \in d} \prod_{w_j \in C(w_i)} p(w_j|w_i) \quad (1)$$

where D is a text corpus of documents, w is a term in a document d and $C(w)$ is the set of terms in w's skip-gram (Zhang et al., 2023). The likelihood is defined using the von Mises-Fisher (vMF) distribution, a commonly employed choice in topic modeling studies (Batmanghelich et al., 2016) (Li et al., 2016), (Jameel and Schockaert, 2019), (Meng et al., 2020):

$$p(w_j|w_i) = \frac{exp(k_{w_i} cos(u_{w_i}, v_{w_j}))}{\sum_{w'} exp(k_{w_i} cos(u_{w_i}, u_{w'}))} \quad (2)$$

where u_{w_i} and v_{w_j} are the embeddings of terms w_i and w_j, respectively, and $k_i \geq 0$ is the concentration parameter that indicates the semantic specificity of w_i.

The objectives for document and category are defined as follows:

$$J_{doc} = \log \prod_{d \in D} \prod_{w \in d} p(d|w) \quad (3)$$

$$J_{cat} = log \prod_{1 \leq i \leq |S|} \prod_{w \in T_i} p(c_i|w) \quad (4)$$

where S is a set of input seeds, T_i is a set of terms related to the seed s_i and c_i is a semantic category represented by s_i. As in the case of skip-grams, likelihoods are defined using the von Mises-Fisher distribution and are computed using v_d and v_c, which represent the embedding vectors of document d and category c, respectively.

The final embedding objective facilitates all three types of contexts. Embedding learning follows optimization process introduced in CatE (Meng et al., 2020):

$$\max J_{emb} = \max \left(J_{skip} + J_{doc} + J_{cat}\right) \quad (5)$$

such that

$$||u_w|| = ||v_w|| = ||v_d|| = ||v_c|| = 1$$

Subsequently, each term w in the training corpus acquires two seed-guided text embedding vectors u_w and v_w, representing semantics of w as the context term and center term, respectively. In the later stages of the algorithm, the semantic closeness between the retrieved term and its seed is computed using the cosine similarity between their respective embedding vectors.

$$sim_{Emb}(w, s_i) = cos(u_w, u_{s_i}) \quad (6)$$

The second context leverages *Pre-trained Language Model Representations*. Models such as BERT acquire extensive general knowledge from vast corpora, such as Wikipedia, which can complement the information within the input corpus.

Given our project's focus on Slovene text data, the use of original BERT models is not feasible for our specific corpora. Therefore, we have the option to employ a pre-trained monolingual Slovene BERT-like model, *SloBERTa*, or a trilingual *CroSloEngual BERT* model, trained on Croatian, Slovenian, and English corpora, offering the potential for cross-lingual knowledge transfer. Additionally, the Spacy library provides a set of pipelines of varying sizes trained on news in Slovenian.

PLM-based corpus-level semantics of a term w are represented by vectors produced by averaging sentence-level PLM representations of w's mentions $w_{1,..,M}$:

$$h_w = \frac{1}{M}\sum_{i=1}^{M} PLM(w^i) \qquad (7)$$

Semantic proximity between PLM-based representations of a term w and seed s_i is calculated using cosine similarity:

$$sim_{PLM}(w, s_i) = cos(h_w, h_{s_i}) \qquad (8)$$

Finally, the framework assesses whether the utilized *Context* information is *Topic-Indicative* or not. For each seed s_i, it assumes the existence of a set of topic-indicative sentences $\Theta_i = \{\theta_{i1}, ..., \theta_{|\Theta_i|}\}$. Assessment of the semantic proximity between a term and a sentence is based on the term's popularity and distinctiveness. A sentence can be defined as topic-indicative if the term appears frequently within it:

$$pop(w, \Theta_i) = log(1 + tf(w, \Theta_i) \qquad (9)$$

$$tf(w, \Theta_i) = \sum_{j=1}^{|\Theta_i|} tf(w, \theta_{ij}) \qquad (10)$$

Furthermore, the term must demonstrate significantly greater relevance to its topic-indicative sentence compared to sentences related to other topics. This level of relevance can be quantified using the BM25 function (Robertson and Walker, 1994).

$$d(w, \Theta_i) = \frac{exp(bm25(w, \Theta_i))}{1 + \sum_{i'=1}^{|S|} exp(bm25(w, \Theta_i'))} \qquad (11)$$

Finally, the semantic proximity between a term w and a semantic category c_i, represented by an input seed s_i, is defined as the product of the popularity and distinctiveness of w in the set of topic-indicative sentences Θ_i associated with s_i:

$$sim_{Sntn}(w, c_i) = pop(w, \Theta_i)^{\alpha} d(w, \Theta_i)^{1-\alpha} \qquad (12)$$

where $0 < \alpha < 1$ is a hyperparameter.

5.3.2 Initial Term Ranking

The framework initiates with a singular seed s_i assigned to each semantic category c_i. Given a training corpus D and a set of seeds S, embedding vectors u_w and v_w are generated for each term w in the corpus, along with their similarity scores $sim_{Emb}(w, s_i)$ with input seeds. Simultaneously, a pre-trained Pre-trained Language Model (PLM) computes representations h_w that are utilized to evaluate another measure of semantic proximity $sim_{PLM}(w, s_i)$.

During the initial iteration of the algorithm, the set of topic-related terms T_i comprises only the seed s_i. In subsequent iterations, it expands to include newly discovered terms. For each semantic category c_i, an initial score $score_{Ini}(w, c_i)$ is computed for each term w:

$$\sum_{t_{ij}\in T_i} sim_{Emb}(w, t_{ij}) \sum_{t_{ij}\in T_i} sim_{PLM}(w, t_{ij}) \qquad (13)$$

Finally, τ terms with the highest semantic proximity score $score_{Ini}(w, c_i)$ are added to the set of topic-indicative terms T_i of category c_i.

5.3.3 Topic-Indicative Sentence Retrieval

Given a set of updated topic-indicative terms T_i, the framework identifies a set of topic-indicative sentences Θ_i from the text corpus. Initially, it retrieves "anchor" sentences, defined as those containing topic-indicative terms exclusively from a single semantic category. Formally, for a category c_i, each sentence θ in D is examined for occurrences of topic-indicative terms T_i and the ones with the largest number of terms are extracted:

$$\max_{\theta\in D} count(\theta, c_i) = \max_{\theta\in D} \sum_{w\in T_i} tf(w, \theta) \qquad (14)$$

Given that the "anchor" sentence should not contain any occurrences of terms indicating other categories, the sentences are filtered and ranked based on the following condition:

$$(\forall j \neq i)\ count(\theta, c_i) = 0$$

In order to adjust the *SeedTopicMine* framework to the needs of *News Linking* project, we not only extract the sentences but also the IDs of the documents that encompass them. Additionally, we capture the sentence's "score", representing the count of the seed's topic-indicative terms contained within the sentence.

Next, the framework searches for the neighbors of the "anchors". They are considered as topic-indicative in case they don't contain terms the categories other than the one associated with the "anchor". More specifically, when provided with an "anchor" sentence θ_{ij}, we examine its surrounding sentences at distances of ± 1, ± 2, ..., $\pm y$ within the document. If the $+k$ (or $-k$) sentence contains topic-indicative terms from categories other than the designated topic, we stop the search in that direction. It's important to note that a neighboring sentence is added to the set even if it doesn't contain topic-indicative terms of the current seed.

The search produces a union of the "anchor" sentence set and their neighbours resulting in a set of topic-indicative sentences Θ_i for each seed s_i.

5.3.4 Ensemble of Multiple Types of Contexts

Finally, the framework ensembles all three types of contexts and combines the corresponding measures of semantic proximity between a term w and a category c_i:

$$score_{All}(w, c_i) = score_{Ini}(w, c_i)\ sim_{Sntn}(w, c_{ij}) \tag{15}$$

The terms are sorted based on the total similarity score $score_{All}(w, c_i)$, and only a limited number of terms with the highest score are passed to the rank ensemble stage.

5.3.5 Retrieval of topic-related documents

In order to adjust the seed-guided topic discovery to the task of document linking, we also count the occurrences the extracted terms within the respective documents. To be specific, for each input seed s_i, each seed-indicative term t_{ij} and each seed-indicative sentence θ_{ik} we count a number of occurrences of t_{ij} in θ_{ik}.

However, rather than storing the term count per topic-indicative sentence θ_{ik}, we opt to save it per the document d_{ik} containing that sentence. This decision is driven by our focus on linking entire documents rather than individual sentences. Furthermore, each document may encompass sentences related to multiple topics, with each of these topics potentially having several topic-indicative sentences within the same document.

The counting procedure is carried out only after acquiring a list of terms supported by all three contexts, even though it could have been conducted during the selection of topic-indicative sentences without the need for additional computations. However, during that stage, the terms were obtained solely from a single context, making them less reliable.

It's crucial to emphasize two key points: (1) we employ a basic term counting procedure to identify topic-related documents, and (2) our searches are confined to a set of topic-indicative sentences obtained in the earlier stage of the training iteration.

An alternative search approach could involve exploring the terms within the entire documents that contain the topic-indicative sentences. However, considering that we extract a substantial number of sentences per each topic (typically 500 by default) and that there could be numerous "anchor" sentences along with their corresponding "neighbors" within this set, we can assume that they comprehensively cover the relevant portion of the document. Conducting an additional document-level search might incur significant time expenses without substantially altering the counts.

5.3.6 Rank Ensembling

All terms are ranked in a descending order based on three types of scores, namely, the final similarity score $score_{All}(w, c_i)$, the score produced by seed-guided text embeddings and the one based on PLM representations. Each term w has three ranks $r_{All}(w|c_i)$, $r_{Emb}(w|c_i)$ and $r_{PLM}(w|c_i)$ that are ensembled into a final Mean Reciprocal Rank $MRR(w|c_i)$:

$$\frac{1}{3}\ (\frac{1}{r_{All}(w|c_i)} + \frac{1}{r_{Emb}(w|c_i)} + \frac{1}{r_{PLM}(w|c_i)}) \tag{16}$$

Lastly, a set of topic-indicative terms is updated with the terms whose MRR score exceeds a certain threshold η.

$$T_i = \{w \mid MRR(w|c_i) \geq \eta\},\ (1 \leq i \leq |S|) \tag{17}$$

The corresponding document IDs for the terms are also saved. While each term encompasses a substantial number of related documents, forming the so-called "extended document list," a typical ground-truth document typically has around 5-15 links to related documents. To narrow down the selection to a small number of the most relevant documents, we once again employ a Mean Reciprocal Rank (MRR) utilizing a two-tiered document ranking system.

The first tier ranks documents based on the frequency of their sentences being chosen as topic-indicative across all seeds and terms. The second tier involves counting the number of seeds for which the document's sentences were chosen as topic-indicative. The rationale behind these two metrics is as follows: since our input document is represented by multiple seeds, the more seeds that identify related sentences in the training document, the higher the probability that it covers the same topic as the testing one.

Likewise, the more topic-related sentences different selected terms have within a training document, the more related the given document is. Therefore, the ranking process generates a list of documents sorted by their Mean Reciprocal Rank (MRR). Preceding the ranking, each term of each seed maintained a list of documents containing topic-indicative sentences where those terms appeared.

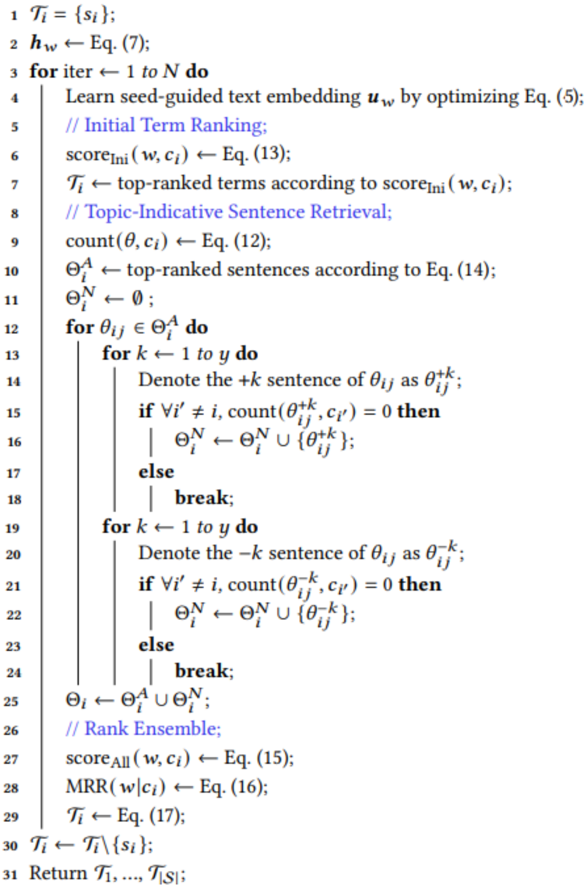
The ranking process consolidates all these lists from various seeds and terms into a unified list of document IDs, providing predictions for documents related to the input testing document.

The updated sets of topic-indicative terms $T_1, ..., T_{|S|}$ are then passed to the next iteration, where the initial term ranking, topic-indicative sentence retrieval, and rank ensemble stages are repeated. The pseudocode is presented in the Figure 4.

5.3.7 Evaluation metrics

Given discovered terms under each seed, we are planning to evaluate the extracted terms based on *topic coherence* that is computed with the help of **NPMI** (Lau et al., 2014) metric. It functions as a frequently employed metric in topic modeling to evaluate the coherence of topics within a given subject. This determination involves computing the average normalized pointwise mutual information across every pair of terms.

$$npmi = \frac{1}{|S|}\sum_{i=1}^{|S|}\frac{1}{\binom{|T_i|}{2}}\sum_{t_{ij},t_{ik}\in T_i}\frac{\log\frac{P(t_{ij},t_{ik})}{P(t_{ij})P(t_{ik})}}{-\log P(t_{ij},t_{ik})} \quad (18)$$



```
$\mathcal{T}_i = \{s_i\}$;
$h_w \leftarrow$ Eq. (7);
for iter $\leftarrow 1$ to $N$ do
    Learn seed-guided text embedding $u_w$ by optimizing Eq. (5);
    // Initial Term Ranking;
    $\text{score}_{\text{Ini}}(w, c_i) \leftarrow$ Eq. (13);
    $\mathcal{T}_i \leftarrow$ top-ranked terms according to $\text{score}_{\text{Ini}}(w, c_i)$;
    // Topic-Indicative Sentence Retrieval;
    $\text{count}(\theta, c_i) \leftarrow$ Eq. (12);
    $\Theta_i^A \leftarrow$ top-ranked sentences according to Eq. (14);
    $\Theta_i^N \leftarrow \emptyset$ ;
    for $\theta_{ij} \in \Theta_i^A$ do
        for $k \leftarrow 1$ to $y$ do
            Denote the $+k$ sentence of $\theta_{ij}$ as $\theta_{ij}^{+k}$;
            if $\forall i' \neq i, \text{count}(\theta_{ij}^{+k}, c_{i'}) = 0$ then
                $\Theta_i^N \leftarrow \Theta_i^N \cup \{\theta_{ij}^{+k}\}$;
            else
                break;
        for $k \leftarrow 1$ to $y$ do
            Denote the $-k$ sentence of $\theta_{ij}$ as $\theta_{ij}^{-k}$;
            if $\forall i' \neq i, \text{count}(\theta_{ij}^{-k}, c_{i'}) = 0$ then
                $\Theta_i^N \leftarrow \Theta_i^N \cup \{\theta_{ij}^{-k}\}$;
            else
                break;
    $\Theta_i \leftarrow \Theta_i^A \cup \Theta_i^N$;
    // Rank Ensemble;
    $\text{score}_{\text{All}}(w, c_i) \leftarrow$ Eq. (15);
    $\text{MRR}(w|c_i) \leftarrow$ Eq. (16);
    $\mathcal{T}_i \leftarrow$ Eq. (17);
$\mathcal{T}_i \leftarrow \mathcal{T}_i \backslash \{s_i\}$;
Return $\mathcal{T}_1, ..., \mathcal{T}_{|S|}$;
```

Figure 4: Pseudocode of the original *Seed-TopicMine* algorithm (Zhang et al., 2023)

We evaluate the task of document linking using the *accuracy* metric. In this evaluation, we consider two equal-sized sets of IDs representing related documents: the ground-truth set D_i and the prediction set $\hat{D}_i$. The accuracy metric is determined by the ratio of correct predictions to the total number of defined related documents. To facilitate this calculation, the number of correct predictions is identified as the size of the intersection set between D_i and $\hat{D}_i$.

$$acc1 = \frac{|Intersection(D_i, \hat{D}_i)|}{|D_i|} \quad (19)$$

Furthermore, we assess the ”extended document list” $D\hat{E}_i$ by examining the count of ground-truth document IDs present within it. It is crucial to compute these two metrics independently to evaluate both the similarity-score-based document retrieval procedure outlined in Section 5.3.5 and the two-tier MRR document ranking described in Section 5.3.6, as the latter may impact the performance of the former.

$$acc2 = \frac{|Intersection(D_i, D\hat{E}_i)|}{|D_i|} \qquad (20)$$

6 Experimental setup

Experimental procedure involves the following steps:

1. Task definition
2. Dataset preparation
3. Train/test data split
4. Preparation of input seeds
5. Preparation of word embeddings
6. Preparation of PLM-based representations
7. Settings and metrics
8. Retrieval of topic-indicative terms
9. Retrieval of topic-indicative documents

6.1 Task Definition

Problem Definition. Given a text corpus of articles, text descriptions of photos and videos $D = \{d_1, ..., d_{|D|}\}$ and an input document d_t such that $d_t \notin D$, the news linking task aims to find a set of documents $D_{tr} = \{d_{tr1}, ..., d_{tr|D_{tr}|}\}$ appearing in D such that each document d_{tr_i} is semantically close or is 'topic-related' to d_t.

To solve this problem we incorporate a modified *SeedTopicMine*(Zhang et al., 2023) seed-guided topic discovery framework. Its goal is given a corpus D and a set of seeds $S = \{s_1, ..., s_{|S|}\}$ to find a set of terms $T_i = \{t_{i1}, ..., t_{i|Ti|}\}$ appearing in D for each seed s_i where the term t_{ij} is semantically close to s_i.

We have extended the functionality of *SeedTopicMine* to not only extract sets of seeds T_i, but also, for each seed t_{ij}, to retrieve a set of documents $D_{ij} = \{D_{ij1}, ..., D_{ij|D_{ij}|}\}$ related to the identified terms. The document sets are aggregated across all terms and seeds, and subsequently, the top-k documents are chosen based on a scoring function. This selection process yields a conclusive set of predictions$D_{tr} = \{d_{tr1}, ..., d_{tr|D_{tr}|}\}$.

6.2 Dataset preparation

We are using a single STA API dataset containing Slovenian articles, photos and videos published by STA during the timeframe from January 1, 2023, to December 11, 2023. Our decision to rely on a single dataset is prompted by the absence of publicly available Slovenian datasets annotated with links to related documents. The limited resources for the Slovenian language leave us with only one viable text corpus, namely the ”SUK 1.0” corpus, comprising 2913 texts. However, it lacks related documents field that essential for our experiment's target, making it unusable for our purposes.

Consequently, we rely solely on documents collected from the STA dataset. During the preprocessing phase, we excluded unnecessary categories such as traffic information, daily bulletins, and condensed daily digests. The remaining documents underwent basic preprocessing procedures to align with the format of the NYT news corpus used in our paper experiments.

STA Articles, photos and videos have slightly different formats, therefore a unification to a common format is done in the preparation Python script. Documents of the resulting dataset contain unique IDs, texts themselves, lists of keywords and lists of IDs of the related documents.

6.3 Train/Test data split

We've collected X documents in total. Our strategy involves dividing this collected data into a training corpus and a compact testing set comprising 50 documents designated for the evaluation phase. The rationale behind opting for such a modest testing size is rooted in the limitations of the framework's execution.

The model operates on an iterative approach, restricting batch inference to only one input document and one set of input seeds at a time. A standard framework inference involves four iterations, with each iteration taking approximately 25 minutes. This results in a total of 100 minutes per document on CPU. The predominant portion of execution time is consumed by the training of seed-guided text embeddings, necessitating the framework to traverse the entire training corpus once for each term of the seed.

Therefore, due to limitations in time and computational power we've decided to stick to a small validation set, as we believe that it still can give relatively representative results. We believe that despite its size, it can still yield relatively representative results. Another influencing factor in our decision is the challenge posed by language translation, as we are not proficient in Slovenian. At least brief examination of the documents is essential to select representative input seeds and to assess the relevance of predicted documents to the input document.

However, only a portion of the collected STA documents includes IDs of related documents. Consequently, we randomly selected 50 samples from this subset of the corpus. Lists of related documents contain a different number of IDs which may harm the consistency of our evaluation. To mitigate this, we made the decision to restrict our evaluation candidates to documents that precisely have 10 related documents. During the evaluation stage, these documents are compared with the 10 predictions of document IDs featuring the highest scores.

As a result, our training corpus consists of X1 documents, X2 words, and a vocabulary of size X3. To streamline this process, we developed a Python script that partitions the initial STA data into a training corpus and a validation set, randomly sampling 50 testing documents in accordance with the aforementioned criteria.

6.4 Preparation of the input seeds

Furthermore, the preparation Python script processes the texts of the documents through a Slovenian Named Entity Recognition (NER) pipeline obtained from the spaCy Python library. The resulting extracted seeds are stored as potential candidates for input seeds to be fed into the main algorithm. Subsequently, we individually examine the test documents to assess whether the extracted named entities accurately represent the document's content. In instances where there is a shortage of seeds or if they fail to cover certain aspects of the text, we supplement them manually with seeds of our choice. However, this manual selection poses challenges, particularly in navigating language barriers.

6.5 Preparation of word embeddings

We rely on the CLARIN.SI-embed.sl collection (Terčon et al., 2023) for our word embeddings. This collection consists of embeddings derived from an extensive set of Slovenian texts, encompassing existing Slovenian corpora. The embeddings are generated using the skip-gram model of fastText, trained on 5,791,405,942 tokens of running text. We store embeddings for the 250,000 most frequently used Slovenian words.

6.6 Preparation of PLM-based representations

Prior to initiating the training process, it is essential to generate PLM-based representations of the training dataset's vocabulary. To accomplish this task, we opted for a pre-trained SloBERTa model available in the HuggingFace library. The SloBERTa model, a monolingual Slovene BERT-like model, shares close ties with the French Camembert model (Martin et al., 2020). Given the scarcity of resources for the Slovenian language, SloBERTa emerged as the sole viable pre-trained Slovenian transformer we could locate. The PLM-based representations of the vocabulary are stored in a dedicated file, crucial for computing PLM-based similarity scores between the seeds and terms.

6.7 Settings and Metrics

- Training corpus: STA dataset (01/23-12/23), X documents

- Testing dataset: 50 randomly selected STA documents with a list of related documents of size 10

- *SeedTopicMine* framework is used to extract topic-indicative terms, with the introduced adjustments facilitating the collection and aggregation of identifiers of the corresponding topic-indicative documents.

- The framework operates on a single document at a time and lacks the capability to perform inference on batches of documents

- For each input document the framework makes 4 iterations over the training corpus

- Each iteration takes roughly 25 minutes on a CPU, so the totak execution time is 100 minutes per input document

- NPMI (Lau et al., 2014) metric is used to assess topic coherence for the topic modeling task

- Accuracy metric is used to compare the ground-truth set of related documents to the set of predicted document links for the news linking task

- The efficiency of document selection and aggregation is assessed by calculating the number of occurrences of ground-truth related documents within the pre-aggregated set of predicted documents

6.8 Retrieval of topic-indicative terms

Method execution follows the procedure described in sections 5.3.1 - 5.3.6 and Figure **??**. Initially, training of seed-guided text embedding is conducted using the official C implementation of CatE (Meng et al., 2020), an embedding learning method. The subsequent steps of the algorithm are implemented in a Python script.

The acquired embeddings are stored in the file system and combined with pre-prepared SloBERTa embeddings to compute initial scores between each term in the corpus and the provided seeds. For each seed, a collection of terms with the highest scores is identified, constituting the topic-indicative term set, which is then saved to the file system as intermediate results.

Given the initial term sets, the framework searches for the topic-inidicative sentences. It is done by counting the number of topic-indicative terms appearing in each sentence and selecting sentences that exclusively contain terms of only a single category as the "anchors". This condition is crucial, since the topic modeling task implies that semantic categories (represented by seeds) should be distinctive.

Next the search continues in the preceeding and suceeding sentences of the anchors. They are added to the set unless they contain terms of the other categories. Consequently, sentences containing terms from multiple seeds are not added neither to the set of "anchors", nor their neighboring sentences. Union of the two sets is obtained and saved to the file system. The size of a set of topic-indicative sentences is limited to 500 samples.

For each term in the training corpus its semantic closeness with each topic-indicative sentences is calculated based on criteria a *popularity* and *distinctiveness*. Popularity criteria implies that a term close to a particular seed would occur frequently in its topic-indicative sentences, while the distinctiveness means a term would be more relevant to the sentences pf the corresponding seed, rather than to sets of sentences belonging to other categories.

In the concluding step, an ensemble is formed by combining scores from three types: the current score, SloBERTa-based scores, and seed-guided embeddings-based scores. The ensembled scores are saved to the file system as the intermediate results. The combined scores contribute to the production of three ranks, which are then further integrated using the Mean Reciprocal Rank (MRR) calculation.

Before computing the MRR, only the terms with the top 20 scores are retained. Terms with a reciprocal rank below the threshold of 0.3 are excluded, while the remaining terms are added to the respective sets of topic-indicative sentences.

6.9 Retrieval of topic-indicative documents

In order to obtain a list of topic-related documents some modification were made to the *SeedTopicMine* algorithm. The first modification is applied at the stage of topic-indicative sentences retrieval, as it captures not only the sentence text, but also the ID of the document it belongs to.

The subsequent modification occurs after computing the scores of the ensembled contexts. At this stage of the algorithm, each seed possesses sets of topic-indicative terms and sentences. Therefore, for each input seed s_i, each seed-indicative term t_{ij} and each seed-indicative sentence θ_{ik} we count a number of occurrences of t_{ij} in θ_{ik}. However, we add the obtained term count to the score of θ_{ik}'s document d_{ik}.

We save this information in a dictionary with the key being the document ID and the value being the number of occurrences of the term within the seed's topic-indicative sentences belonging to the document. Thus for each term of the seed we obtain a list of documents and the corresponding term counts. We save the dictionary to the file system.

The last modification to the algorithm happens at the rank ensemble stage. Since we obtain an updated set of topic-indicative terms, we transfer the counts from the dictionary obtained in the previous stage to the corresponding terms and save the updated dictionary.

The original *SeedTopicMine* algorithm concludes at this point, but we have introduced an additional stage that consolidates counts across different terms and seeds. The objective is to create a dictionary where the key represents the document ID, and the corresponding value indicates the frequency with which the document's sentences were selected as topic-indicative for any seed or term. We can securely aggregate these term counts across all seeds, because we treat each seed as a representative of the input document.

Moreover, we count the number of seeds for which the document's sentences were chosen as topic-indicative. These two scores are utilized to generate two corresponding ranks, and the Mean Reciprocal Rank (MRR) is computed to organize the list of document IDs. This expanded document list typically contains 50-100 documents, whereas the ground-truth lists typically range from 5-10 documents. We simply select the corresponding number of the top-ranked documents to produce the final prediction. We calculate two accuracy metrics, one involving the full document list, the other using the reduced one.

7 Results

We conducted four iterations of the modified *SeedTopicMine* algorithm for each of the 50 test documents, utilizing ground-truth lists containing 10 related documents. For every input document, we generated a list of predicted IDs for the related documents. The "extended" lists consisting of more than 30 documents were reduced to top-10 documents based on their scores. As outlined in Section 5.3.7, both the original and modified lists were employed to calculate accuracy metrics, namely "Accuracy 2" and "Accuracy 1", respectively. The metrics were averaged over the 50 input samples.

Additionally, we computed NPMI metrics for the terms of each input seed for every input document. Subsequently, we averaged the NPMI scores across the document's seeds, resulting in one score per document. Finally, these scores were aggregated across all 50 testing samples. The evaluation results are presented below:

- **Accuracy 1**: 0.0067
- **Accuracy 2**: 0.2938
- **NPMI score**: 0.4215

We've received unsatisfactory results for the news linking task and slightly better ones for the task of topic modeling. The accuracy calculated on prediction sets of size 10 is close to 0 and for only 3 documents we had non-zero intersection of the ground-truth and prediction sets.

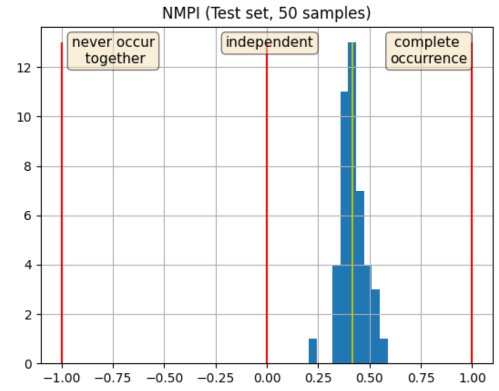


Figure 5: Histogram of NPMI score computed on the testing set

Slightly better results were obtained when computing intersection with the "extended" document list. We can conclude that at least 1/3 of the desired documents are extracted by the modified algorithm, but are lost at the document list aggregation stage.

NPMI is a metric that yields scores close to -1 when terms never occur together, 0 when terms are independent, and 1 when they frequently cooccur. As illustrated in Figure 5, the resulting NPMI scores fall approximately midway between 0 and 1. Since the STA dataset has not been tested with any other topic modeling methods, it is challenging to definitively assess the performance of *SeedTopicMine*.

Nevertheless, the paper's experiments showcase the framework's state-of-the-art (SOTA) NPMI results on NYT-Topic and NYT-Location benchmarks, achieving scores of 0.1947 and 0.2176, respectively. With an NPMI score twice as large on STA dataset, we may assume that the at least half of the extracted terms is properly related to the respective seeds.

In the subsequent section, we discuss the unsatisfactory performance of the framework in the context of the news linking task.

8 Discussion

8.1 Distinctive topics

SeedTopicMine assumes that each input seed represents a distinct semantic category, meaning that each category c_i is defined by a single seed s_i. Consequently, when presented with a set of N input seeds, the framework treats them as unique representatives of N semantic categories. semantic categories. This becomes pivotal during the selection of topic-indicative sentences. Given a seed s_i and its associated set of terms T_i, the algorithm seeks sentences containing terms from T_i, but devoid of any topic-indicative terms from other seeds.

For example, consider a scenario where the user inputs seeds such as *"Metallica"* and *"Warsaw"* to find reviews on Metallica's concerts in Warsaw. In this case, for the seed *"Warsaw"*, the algorithm searches for sentences without terms related to *"Metallica"* and vice versa. Consequently, sentences containing both *"Warsaw"* and *"Metallica"* are excluded, leading to non-inclusion of documents lying in the user's topic of interest.

The algorithm identifies sentences containing terms exclusive to the given seed and not used for other seeds. This behavior was observed in an analysis of results generated for an article discussing the election of a new Speaker of the American Congress, Mike Johnson from the Republican Party, replacing Kevin McCarthy. The input seeds included *"Johnson," "McCarthy," "Kongres"* (Congress), and *"Republikanec"* (Republican). The framework did not output any documents containing both *"Johnson"* and *"McCarthy"* or *"Johnson"* and *"Kongres"*. Instead, it produced documents describing other individuals with the surname Johnson, unrelated to the Congress and Republican Party.

We've tested the framework by passing it a set of seeds in such a way that all of them are tied to a single semantic category. However, we observed that the algorithm recognized only the first word as a seed. The remaining words were considered as topic-related terms and were not utilized in the initial term ranking. Consequently, they did not influence the selection of the updated set of topic-indicative terms. Since they were not related to the initially selected seed, their scores were low, and they were subsequently removed from the set after a single iteration of the algorithm.

Recognizing this issue, we are planning to enhance the framework by enabling it to represent each semantic category with a *set* of seeds, rather than a single word. Consequently, each semantic category c_i will be represented by a set of input seeds $S_i = \{s_{i1}, s_{i2}, ..., s_{i|S_i|}\}$. Implementing this modification would necessitate changes in different parts of the algorithm, particularly in the calculations of the similarity score for seed-guided text embeddings (Equation 6) and PLM-based representations (Equation 8), which would need to be adjusted as follows:

$$sim_{Emb}(w, c_i) = \sum_{s_{ij} \in S_i} cos(u_w, u_{s_{ij}}) \tag{21}$$

$$sim_{PLM}(w, c_i) = \sum_{s_{ij} \in S_i} cos(h_w, h_{s_{ij}}) \tag{22}$$

While the calculation of PLM-based representations is seed-indifferent, the learning of seed-guided text embeddings is conducted with the assistance of an objective that relies on the representation of the semantic category c_i. Specifically, the term $p(c_i|w)$ from an Equation 4 is calculated as:

$$p(c_i|w) = \frac{exp(k_w cos(u_w, v_c))}{\sum_{c'} exp(k_w cos(u_w, v_{c'}))} \tag{23}$$

Currently, it's unclear whether the learning of the embedding v_{c_i} of the category c_i would be influenced by an expansion of the seed set S_i. This aspect needs further examination within CatE's C-language implementation of the embedding learning process. Assuming that the learning process remains unaffected by the change, the selection of the set of topic-indicative terms is determined based on the score (Equation 13), which incorporates the similarities between each word in the training corpus and the *set* of seeds, rather than a single seed.

As a result, words related to more than one category's seed will be prioritized, leading to a majority of extracted terms being associated with a combination of category seeds. Importantly, since the selection of topic-indicative sentences depends solely on the category's set of topic-indicative terms, no alterations need to be made to this aspect of the algorithm.

In summary, our experiments have revealed that the *SeedTopicMine* framework operates in a manner contrary to the objectives of the *News Linking* project. A change in the "seed"-"semantic category" relation has to adjust the algorithm to the requirements of the news linking task.

8.2 NER and choice of the seeds

While the final selection of terms was unsatisfactory, as described in the previous subsection, we have also concluded that particular attention must be given to the choice of input seeds. Our program allows users to rely on the results of the NER model and/or manually fill them. Despite utilizing SpaCy's NER pipeline, specifically pre-trained on a set of Slovenian news, it occasionally fails to identify obvious named entities like people or institutions.

Furthermore, the popular choices of named entities are sometimes insufficient to precisely characterize input documents, requiring additional keywords such as dates, events or some domain-specific terms. These elements are often not extracted by the NER model. However, since the STA dataset lacks NER annotations, we cannot use it for fine-tuning the NER model. None of our team members speak Slovenian, which means we must consider the possibility that the input seeds for the document might not be optimal. This, in turn, could be one of the reasons the algorithm performed poorly for the task of news linking.

8.3 Choice of the evaluation metrics

We employed an iterative framework that deviates from typical deep learning solutions, as it doesn't use backpropagation and lacks a distinct separation between training and testing stages. Traditional training, where the network's weights are adjusted according to errors on the training dataset, is performed in this approach. Although the framework encompasses the training of text embeddings, its dependency on input seeds and, in subsequent iterations, on extracted topic-indicative terms makes it unsuitable for application to other seeds and input documents.

Moreover, it's highly possible that additional documents related to a given article exist beyond those provided in the 'related' field of the STA dataset. Consequently, the framework may predict a document describing the same topic but not included in the ground-truth set of document IDs. While this prediction might be semantically accurate, it would reduce accuracy score. We encountered this scenario with an example detailing the change of the Speaker of the USA's Congress from McCarthy to Johnson. Out of the 10 ground-truth IDs of the related documents in the dataset:

- 1 linked to an article describing McCarthy's struggles to retain his position
- 2 linked to articles informing about McCarthy's resignation
- 7 linked to photos annotated with the same text about Johnson becoming Speaker

In contrast, the framework output only 1 document about Johnson, specifically describing his initial actions as the elected Speaker. 4 documents were unrelated to the Speaker elections topic, and the remaining 5 described McCarthy's struggles on keeping his position. While the incorrect choice of the former 5 articles is directly linked to the issue described in Subsection 8.1, each of the 5 articles on McCarthy is semantically close to the one present in the ground-truth set.

Given these issues, we are uncertain about whether we can employ the accuracy metric in the same manner as typical machine learning tasks with a defined set of ground-truth labels.

8.4 Documents retrieval strategy

The reliance on the count of topic-indicative terms within documents as the primary criterion for document selection may introduce bias and inaccuracy. This approach favors topic-indicative sentences, and consequently, the corresponding documents, that contain terms well-established in a large number of documents. Once again, the Johnson-McCarthy example serves as an illustration. McCarthy served as a Speaker for several years and was frequently featured in the news prior to his resignation in late October, when he was replaced by Johnson. As a result, Johnson's presence in the text corpus is limited to articles from only the last two months of 2023. Consequently, the number of documents including the topic-indicative term *'McCarthy'* is significantly larger than those including *'Johnson'*.

Under this scoring function, the former terms have a considerably higher probability of being selected compared to the latter, as evident in the output example presented in the previous subsection. Importantly, this issue is distinct from the one involving semantic categories and seeds (Section 8.1). The algorithm did not assign high scores to documents containing terms exclusively related to the seed *'Johnson'*, as they occur less frequently than terms related to a well-known American politician who appears frequently in the press.

Consequently, the choice of the document retrieval strategy needs to be addressed independently, and a more sophisticated method would be desirable to mitigate these biases and inaccuracies, probably including another pass over the text corpus with additional document-level analysis, as opposed to the sentence-level search presented in the original *SeedTopicMine*.

9 Conclusion

To sum up, this project incorporated a seed-guided topic modeling framework *SeedTopicMine* for the task of document linking based on semantic similarity. The adjustments made to the algorithm enabled the retrieval of document IDs that are semantically close to the topic-indicative terms extracted by the original framework. To narrow down the set of predicted documents, a simple scoring function based on frequency count was introduced.

Additionally, we gathered and preprocessed articles, photos, and videos collected from the STA API spanning January 2023 to December 2023. By filtering out technical samples and unifying the format for all three types of media, we successfully formed a unified test corpus.

Testing was conducted on a set of 50 documents, resulting in very low accuracy scores for the document linking task and moderate values of NPMI scores for the topic modeling task. We identified several problems that may negatively impact the framework's performance, proposed a solution for one of the issues, and discussed whether the selected metrics and document retrieval strategy are unbiased and suitable for the given task.

References

[Batmanghelich et al.2016] Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Sam Gershman. 2016. Nonparametric spherical topic modeling with word embeddings. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2016, page 537. NIH Public Access.

[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*.

[Jameel and Schockaert2019] Shoaib Jameel and Steven Schockaert. 2019. Word and document embedding with vmf-mixture priors on context word vectors. ACL.

[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.

[Li et al.2016] Ximing Li, Jinjin Chi, Changchun Li, Jihong Ouyang, and Bo Fu. 2016. Integrating topic modeling with word embeddings by mixtures of vmfs. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 151–160.

[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*.

[Martin et al.2020] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[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.

[Robertson and Walker1994] Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.

[Terčon et al.2023] Luka Terčon, Nikola Ljubešić, and Tomaž Erjavec. 2023. Word embeddings CLARIN.SI-embed.sl 2.0. Slovenian language resource repository CLARIN.SI.

[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.