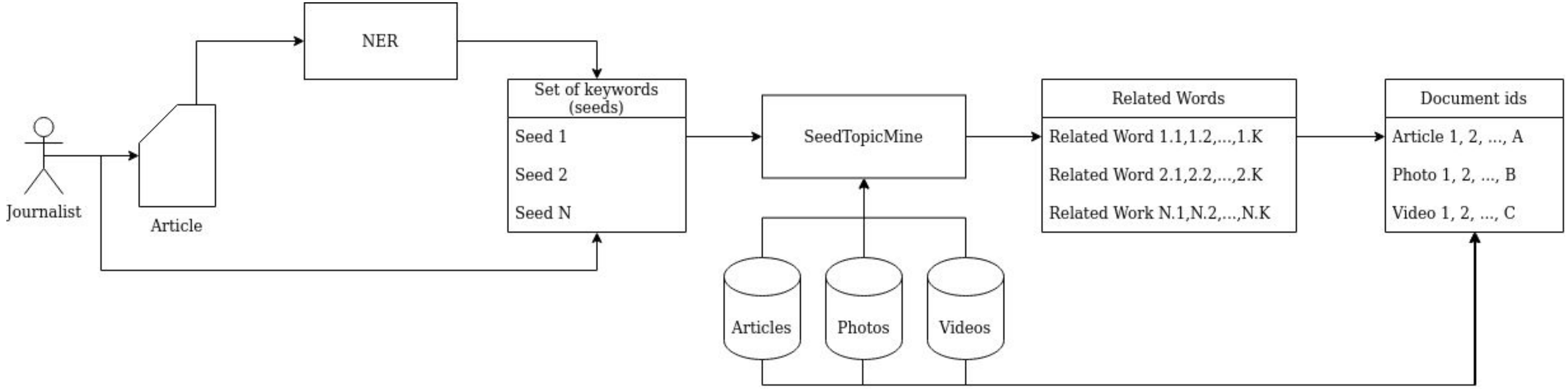
News Linking

Illia Tesliuk
Panpan Liu
Trifebi Shina

Goals

- Collect data from API of The Slovenian Press Agency (STA) and form a text corpus
- Prepare a system that would could connect different types of information in Slovenian based on their content/descriptions:
  - News Articles
  - Photos - text descriptions
  - Videos - text descriptions
- Adjust the topic modeling SeedTopicMine framework to the task of documents linking
- Adjust the SeedTopicMine to Slovenian language

System Review



- Keywords are entered manually or extracted by NER model
- Set of seeds is fed to SeedTopicMine framework

- SeedTopicMine extracts words and sentences related to each seed
- We extract documents that contain these sentences, score and sort them

Document retrieval procedure

1. Obtain a list of topic-related terms (with SeedTopicMine)
2. Obtain a list of topic-related sentences that contain these terms (with SeedTopicMine)
3. Gather IDs of the documents that contain these sentences
4. Sort document list by number of their occurrences
5. Select top-10 documents which occurred most frequently
6. Compare with the set of 10 ground-truth related documents

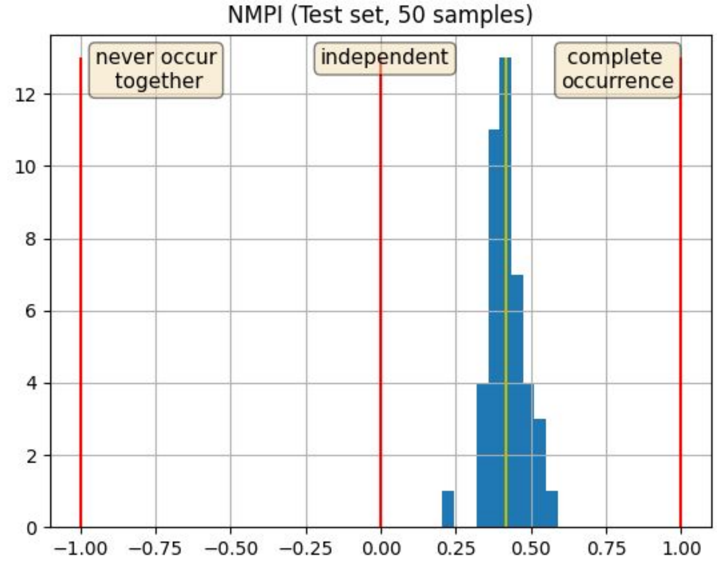
Metrics

- Topic modelling:
  - NPMI - metric that assesses topic coherence and yields scores close to -1 when terms never occur together, 0 when terms are independent, and 1 when they frequently co-occur.
- Document linking:
  - $$acc1 = \frac{|D_i \cap \hat{D}_i|}{|D_i|}$$
  - $$acc2 = \frac{|D_i, \cap \widehat{DE}_i|}{|D_i|}$$

Results

Run	NPMI	Accuracy 1	Accuracy 2
Before seed change (50 samples)	0.4215	0.0067	0.2938
After seed change (15 samples)	0.5361	0.0103	0.3207

Fig. NPMI distribution over the test set for the first experiment



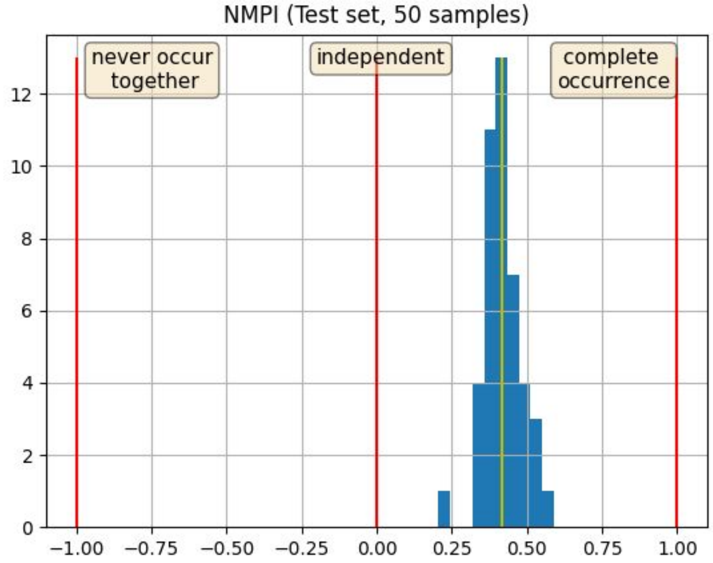
Limitations: Seed-Semantic Category problem

- Framework assumed that each of N input seeds represented a separate semantic category
- Hugely influenced selection of topic-related sentences - they should contain terms related to only a single category
- Example, given the seeds “*Barcelona*”, “*Real Madrid*”, “*Saudi Arabia*”, the original framework wouldn’t include sentences that contain at least 2 of the seeds
- “***Real Madrid** triumphed over **Barcelona** in the Spanish Super Cup final held in **Saudi Arabia***” wouldn’t be selected
- We’ve managed to modify the framework to accept multiple seeds / category

Results

Run	NPMI	Accuracy 1	Accuracy 2
Before seed change (50 samples)	0.4215	0.0067	0.2938
After seed change (15 samples)	0.5361	0.0103	0.3207

Fig. NPMI distribution over the test set for the first experiment



Limitations: Computational power

- Limited computational resources - only private PCs & laptops
- 1 framework iteration ~25 min, 4 iterations/sample = ~100 min
- SeedTopicMine doesn’t support batches
- We had to limit the test set to only 50 documents

Limitations: Language barrier

- By default, seeds are extracted with spaCy NER pipeline
- NER extracted only general named entities, without keywords that uniquely describe the article
- Choice of the seeds heavily influences the final performance
- Manual check by a Slovenian native speaker would have been beneficial

Conclusions

- Framework extracts coherent topic-related terms
- Retrieved documents generally describe topics that are related to the input document, but are not specific enough
- **Example**: given seeds describing election of a new Speaker of Congress, the framework returned documents describing events that led to resignation of the previous Speaker
- Document retrieval method (based on occurrence count) appeared to be biased
- Retrieval method prefers past events as number of related documents is much higher than of the most recent ones

Conclusions

- SeedTopicMine is primarily targeted at topic modeling and operates on a sentence-level analysis
- Document retrieval task uses results of the topic modeling and is secondary to the framework
- SeedTopicMine is not a good choice for the task of documents linking

Potential improvements / alternatives:

- SeedTopicMine can be modified to follow document-level approach in both terms selection and document selection stages
- News linking task somehow resembles recommendation task
- Combination of NLP methods and recommender systems may be more promising than the current implementation

Contributions

- Research on available Slovenian NLP models, datasets, benchmarks and word embeddings
- Gathering a dataset of partially labelled 155,577 articles, photos and videos descriptions from the STA API for the task of document linking
- Modifying and adjusting the topic modelling *SeedTopicMine* framework to the task of the document linking
- Combining *SeedTopicMine* framework with the pretrained Slovenian PLMs for the tasks of NER and word embedding calculations
- Evaluating model performance on a test dataset of size 50 for the tasks of topic modelling and related documents linking

Thank you for attention