News Linker

Illia Tesliuk, Panpan Liu, Trifebi Shina Sabrila

NLP Project 2023

The Evolution of News Reporting

- Growth in diverse media formats beyond text articles


- Challenges of fragmentation
  - Inefficiencies in tracking news developments
  - Disjointed experiences for audiences

Introducing “News Linkers”

- What’s news linkers?
- Unique approach to linking news data
- Move beyond linking to external sources

Research goals

- 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 to solve the news linking problem
- working with low-resource language, adapting high-resource-based model for these needs
- combining different NER, EL and Seed-guided topic modeling into one working solution

Literature Review

Multilingual Text Matching

TF-IDF (Martineau and Finin, 2009)

Word2Vec (Mikolov et al., 2013)

Edit Distance

machine learning model
(e.g., logistic regression)

a statistical measure
(e.g., cosine similarity)

Literature Review

Pre-trained English Model

Transformer (Vaswani et al., 2017)

Bidirectional Encoder Representations from Transformers (BERT) architecture (Devlin et al., 2018)

RoBERTa (Liu et al., 2019)

a task-specific output layer

fine-tune it

Literature Review

Pre-trained Multilingual Model

Multilingual BERT (Pireset al., 2019)	adapt the training strategy from BERT, using a single model's weights to handle all target languages
XLM model (Lample and Conneau, 2019)	parallel corpus sentences that are translations of each other
XLM-R model (Conneau et al., 2019)	balanced the high-resource and low-resource languages in the training data

Slovenian Language

- Corpus: **SUK 1.0** (~1M words)
  - 200k + 340k words with named entity annotations
- Pre-trained models:
  - [Hugging Face] **SloBERTa** - monolingual Slovene BERT-like model
  - [Hugging Face] **CroSloEngual BERT** - trilingual model, trained on Croatian, Slovenian, English corpora; offers cross-lingual knowledge transfer
  - [Clarin.si] **SloNER 1.0** - named entity recognition model trained on SUK1.0 corpus

STA API

- One can retrieve an IDs list of the articles published on a specific date
- Article ID is used to retrieve the content and metadata of a news note in a separate request
- Images and Video Albums can also be retrieved by its ID


```
[ 3198280, 3198550, 3198555, 3198554
3198600, 3198602, 3198603, 3198361,
3198625, 3198631, 3198630, 3198626,
3198652, 3198653, 3198650, 3198662,
3198698, 3198696, 3198671, 3198697,
3198728, 3198726, 3198721, 3198739,
3198776, 3198751, 3198775, 3198774,
3198806, 3198815, 3198812, 3198814,
3198848, 3198849, 3198850, 3198851,
3198777, 3198887, 3198890, 3198891,
```

STA API: Article


```
{
  "byline" : "gj/np/np",
  "channels" : [ "STA" ],
  "desk" : "SP",
  "headline" : "Arca najhitrejša v tržaškem brezvetrju (dopolnjeno)",
  "keywords" : [ "JADRANJE", "REGATA" ],
  "categories" : [ "SP" ],
  "lede" : "Italijanska jadrnica Arca je zmagovalka 55. barkovljanke, najbolj množične regate na svetu, ki je danes potekala v Tržaškem zalivu. Na startu je bilo prijavljenih 1773 jadrnic vseh velikosti, regata pa se je v brezvetrju odvijala po polžje. Najboljša slovenska predstavnica je bila na tretjem mestu Way of Life.",
  "places" : [ {
    "city" : "Trst",
    "country" : "ITALIJA",
    "code1" : "ITA",
    "code2" : "it"
  } ],
  "previous" : 3222411,
  "priority" : 3,
  "id" : 3222446,
  "photos" : [ 1280343, 1280345, 1280346, 1280347, 1280348, 1280349, 1280350, 1280351, 1280352, 1280353, 1280354, 1280355, 1280356, 1280357, 1280358, 1280359, 1280360, 1280361, 1280362, 1280363, 1280364, 1280365, 1280366, 1280367, 1280368, 1280369, 1280370, 1280371, 1280372, 1280373, 1280374, 1280375, 1280376, 1280377, 1280378, 1280379, 1280380, 1280381, 1280382, 1280383, 1280384, 1280385, 1280386, 1280387, 1280388, 1280389, 1280390, 1280391, 1280392, 1280393, 1280394, 1280395, 1280396, 1280397, 1280398 ],
  "videoAlbums" : [ 4343, 4344, 4345, 4346 ],
  "text" : "Pomanjkanje vetra je krojilo današnje dogajanje v Tržaškem zalivu. Hitrost vetra po podatkih specializirane spletne strani Windfinder.com za Barkovlje ob startu regate ni presegala dveh vozlov na uro (3,6 km/h), kar je močno otežilo delo jadralcev na startu.\n\nSapica se je malce okrepila in dosegla štiri vozle po poldnevu in vodilne vendarle potisnila v močno skrajšan cilj regate. Ta je bil po odločitvi žirije že na prvi boji, kar se je po poročanju italijanske agencije Ansa zgodilo prvič v zgodovini.\n\nV zahtevnih jadralskih razmerah se je na koncu najbolje znašla italijanska maksi jadrnica pod taktirko Furia Benussija, ki se je po letu dni premora vrnila na prestol Barcolane. Lani je namreč slavila ameriška jadrnica Deep Blue, ki pa danes ni branila zmage.\n\nV zaključku regate je Benussi krmilo predal 16-letni hčerki Marti Benussi, Arca pa je ciljno črto prečkala v času 1 ure 49 minut in 55 sekund.\n\nNekoliko v ozadju se je odvijal boj med italijansko jadrnico Prosecco Doc (nekdandjim Portopiccolom) pod vodstvom slovenskega krmarja Mitje Kosmine in plovilom Way of Life slovenskega projekta EWOL. Kot druga je z manj kot dolžino prednosti ciljno črto prečkala italijanska jadrnica, ki je za zmagovalko zaostala tri minute.\n\nDanašanja zmaga je sicer druga za Arco na barkovljanki po prvencu leta 2021, ko je bila hitrejša od ekipe Way of Life. Slednja je sicer pod istim imenom, a z drugo jadrnico, zmagala leta 2019.\n\nTudi tokrat je slovenski posadki ob mirnih razmerah kazalo dobro, saj je po startu prišla v vodstvo, ki ga je držala več kot pol poti do skrajšanega cilja. A v drugem delu in ob nekoliko močnejšem vetru je na svoj račun prišla Arca.\n\n\"Morda je ekipa na koncu naredila eno taktično napako, ne glede na to so se borili in bili v vodstvu večji del regate. Porazi so del življenja, iz njih se je treba učiti,\" je o razpletu dejal nekdanji slovenski olimpijski jadralec Gašper Vinčec, sicer vodja projekta EWOL.\n\nVinčec sicer danes ni bil za krmilom najhitrejše slovenske jadrnice, saj je tega tokrat predal Mauriziu Benčiču. Namesto tega pa je sam bil za krmilom
```

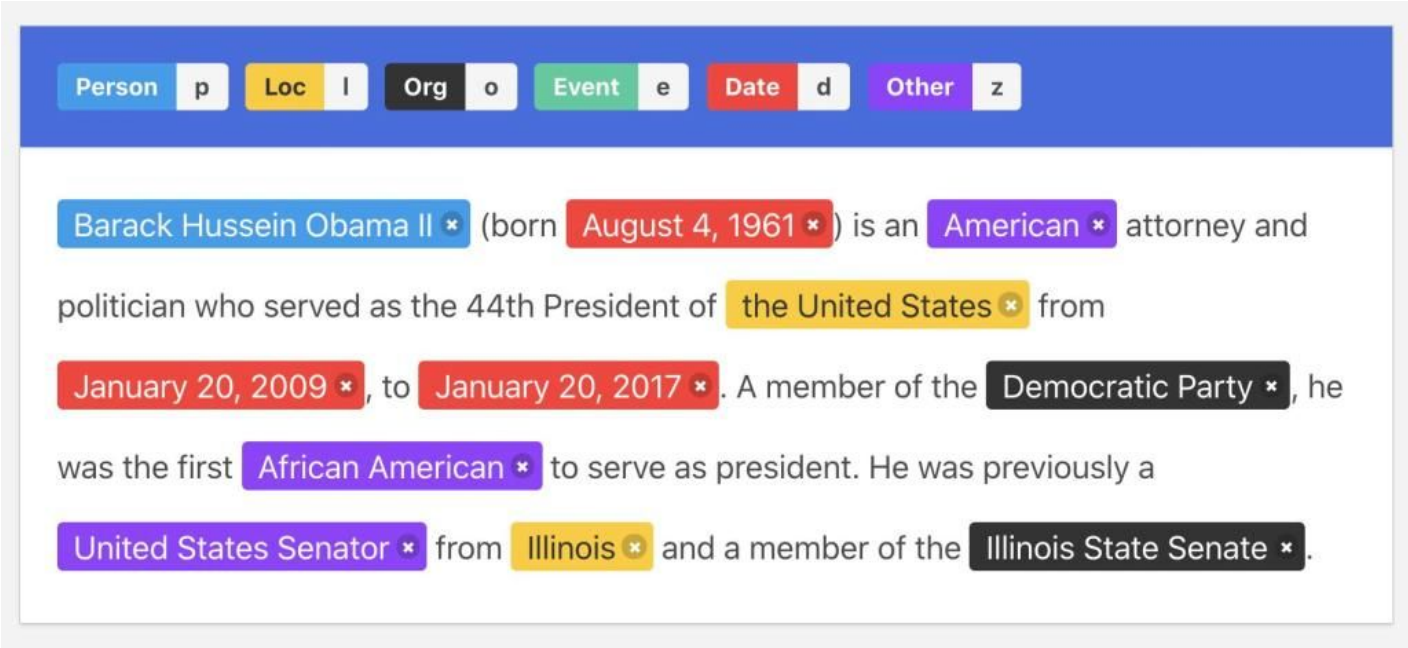
STA API: Image


```
{
  "attachedToArticles" : [ 3200267, 3203100, 3203122, 3207444, 3210357, 3213973, 3221690, 3229243 ],
  "categories" : [ "DR", "KR", "TF" ],
  "tags" : [ "kosovni material", "kosovni odpad", "kosovni odvoz", "posledice poplav", "smeti", "škoda" ],
  "persons" : [ ],
  "created" : 1691998049000,
  "published" : 1692001267007,
  "description" : "Sneberje.\nProstovoljci ob dnevu solidarnosti v pomoč prizadetim v ujmi.\nFoto: STA",
  "enDescription" : "Sneberje\nVolunteers helping in the flood relief effort as part of Solidarity Day, a work-free day.\nPhoto: STA",
  "free" : false,
  "pub" : true,
  "id" : 1263846,
  "albumId" : 115889,
  "agencyId" : 1,
  "width" : 4000,
  "height" : 2252,
  "slAdditionalDesc" : {
    "2023-09-18" : "Sneberje.\nProstovoljci v pomoč prizadetim v ujmi.\nFoto: STA\nArhiv STA"
  },
  "enAdditionalDesc" : {
    "2023-09-05" : "Sneberje\nThe flood damage in Snebrje on the eastern outskirts of Ljubljana.\nPhoto: STA\nFile photo"
  }
}
```

STA API: Video Album


```
{
  "attachedToArticles" : [ 3221192, 3222446 ],
  "date" : "2023-10-08",
  "description" : "55. jadralna regata barkovljanka - priprave na začetek regate",
  "id" : 4343,
  "videos" : [ {
    "duration" : 4.54,
    "created" : 1696754416000,
    "published" : 1696754333000,
    "id" : 16070,
    "width" : 1920,
    "heigth" : 1080,
    "sound" : true,
    "title" : "Ekipa EWOL junior",
    "description" : "55. jadralna regata Barcolana (Barkovljanka)."
  }, {
    "duration" : 13.24,
    "created" : 1696754416000,
    "published" : 1696754416000,
    "id" : 16071,
    "width" : 1920,
    "heigth" : 1080,
    "sound" : true,
    "title" : "Ekipa EWOL junior",
    "description" : "55. jadralna regata Barcolana (Barkovljanka)."
  }, {
    "duration" : 8.08,
    "created" : 1696754416000,
    "published" : 1696767046000,
    "id" : 16074
```

Named Entity Recognition (NER)



Named Entity Recognition (NER)

- task of identifying and categorizing key information (entities) in test
- *named* entity: a word or a series of words that *consistently* refers to the same thing
- every detected entity is classified into a predetermined category
- Categories examples: Person, Place, Date, Organization, Work of Art, etc.
- typical evaluation measures: F1-score, precision, recall

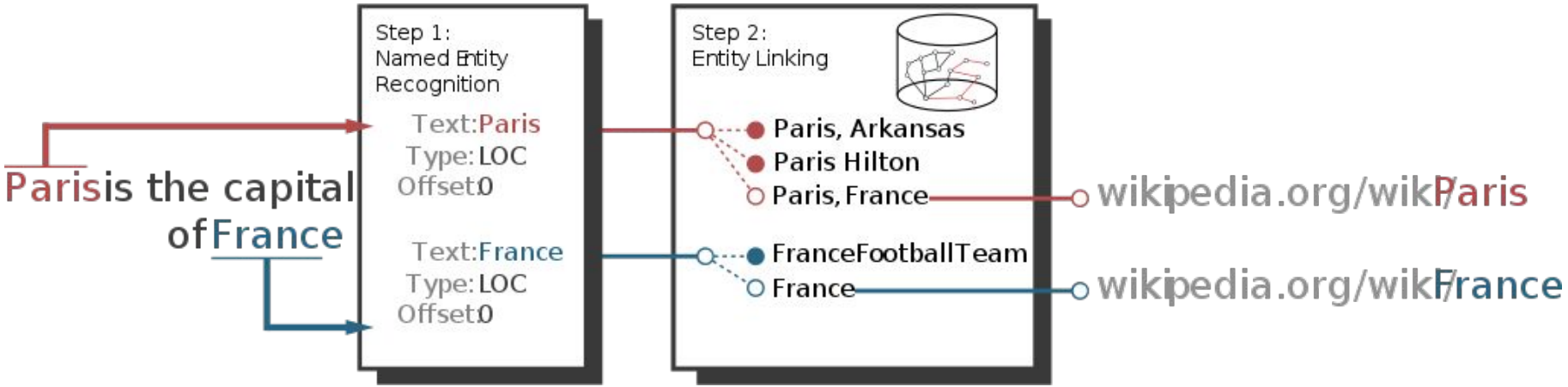
Named Entity Recognition (NER)

- **ACE**: Automated Concatenation of Embeddings for Structured Predictions (Wang et al., 2021)
  f1=***94.6%*** (CoNLL 2003)
  - Automates process of finding concatenations of embeddings
  - A controller alternately samples a concatenation of embeddings and updates the belief based on a reward.
  - Reward is based on the accuracy of a task model

Named Entity Recognition (NER)

- **LUKE**: Deep Contextualized Entity Representations with Entity-aware Self-Attention (Yamada et al., 2020)
  f1=94.3% (CoNLL 2003)
  - proposes new pretrained contextualized representations of words and entities based on the bidirectional transformer.
  - proposes an entity-aware self-attention mechanism that considers the types of tokens (words or entities) when computing attention scores.

Entity Linking (EL)



Entity Linking

- task of **recognizing** (Named Entity Recognition) and **disambiguating** (Named Entity Disambiguation) named entities to a knowledge base (eg. Wikidata, DBpedia)
- There are two approaches:
  - ***End-to-end***: NER + NED
  - ***Disambiguation-only***: directly takes gold standard named entities as input and only disambiguates them to a correct entry in a given knowledge base

Entity Linking

- *Metrics*: micro-precision - fraction of correctly disambiguated named entities in the full corpus
- Evaluating the impact of Knowledge Graph Context on NED Models (Mulang' et al., 2020), M-P=**94.94%**:
  - Authors propose to feed the context derived from a knowledge graph into transformer architectures
  - Additional KG context improves their performance for named entity disambiguation (NED).

Entity Linking

- **DeepType**: Multilingual Entity Linking by Neural System Evolution (Raiman ey al., 2018), M-P=94.88%:
  - Explicitly integrates symbolic information into the reasoning process of a neural network with a type system.
  - Authors construct a type system to constrain the outputs of a neural network to respect the symbolic structure
  - Outperforms solutions that rely on a human-designed type system or recent deep learning-based entity embeddings,

NER for Entity Linking

- Use of NER to extract metadata from news data
- Enhancing content discoverability
- Pretrained models for entity linking

SeedTopicMine

- An iterative seed-guided topic discovery method
- Goal: given a seed produce a set of the corresponding topic-indicative terms
- Example: “business” -> [“firms”,”companies”,”corporations”]
- Jointly learns from three types of context and merges them via ranking ensemble process (Seed-Guided Text Embeddings, PLM representations, Topic-indicative context)

SeedTopicMine

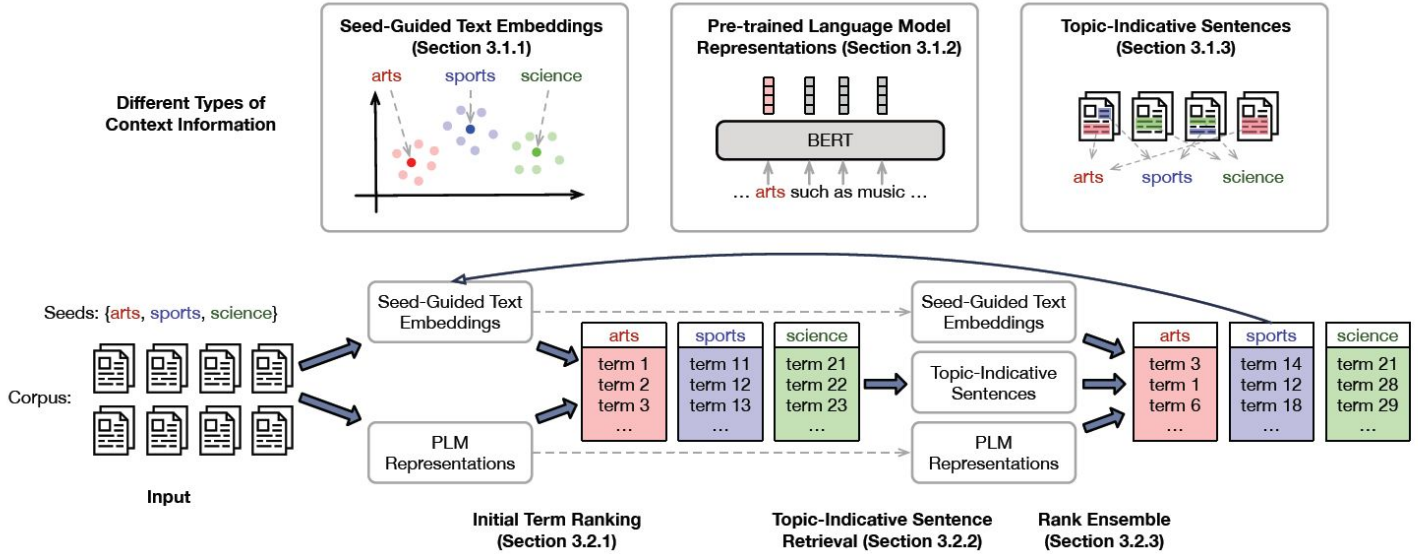


Figure 1: Overview of the SeedTopicMine framework.

Proposed solution

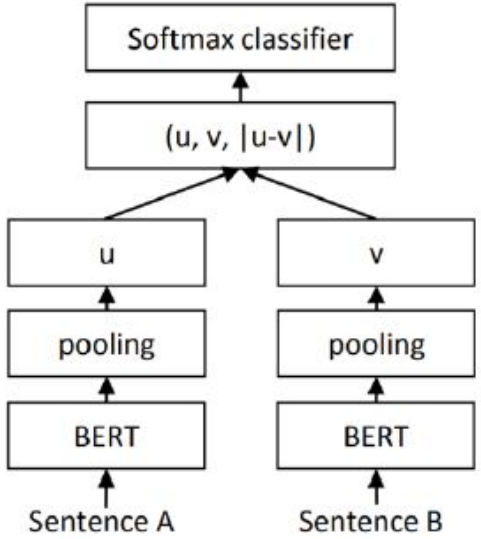
- Gather STA API over a specific period (e.g. 1-3 years)
- Perform preliminary exploratory analysis
- Filter out redundant categories (news digests, etc.) and redundant
- Combine articles, images and video text descriptions in a single text corpus

Approach 1

- Use pretrained **SloNER 1.0** model to retrieve named entities from an input article
- Apply one of the SOTA Entity Linking model to disambiguate named entities
- Feed the retrieved entities as seeds into an adjusted **SeedTopicMine** method that returns seed-related terms together with IDs of the their documents

Approach 2

Use Pre-trained Multilingual Model to Compare Embeddings



Thank you!