EL-NEL: Entity Linking for Greek News Articles

Katerina Papantoniou, Vasilis Efthymiou and Giorgos Flouris

Institute of Computer Science, FORTH, Greece

Department of Computer Science, University of Crete, Greece

Objectives

Named entity linking tool for the Greek language.

- Enable journalists and readers to verify and validate information and opinions without access to the relevant factual resources.
- Trained models are offered as open-source
- Proposed approach applicable to other languages.

DebateLab

The DebateLab project is conducting research towards representing, mining and reasoning with online arguments. The goal of DebateLab is to offer a suite of tools and services that will assist both the work of the professional journalist in accomplishing everyday tasks (e.g., writing, archiving, retrieving articles), as well as the activity of the ordinary Web user (reader) who wished to be well-informed about topics or entities of interest (e.g., persons, locations, events).

Named Entity Linking

An essential component of the envisioned DebateLab tools is the task of identifying sub-strings (aka *entity mentions*) that refer to real-world entities within an unstructured/textual argument, and mapping them to uniquely identified entity descriptions provided in target *Knowledge Bases (KBs)*, a task known as *named entity linking (NEL)*. NEL would allow the association of arguments with factual resources, thereby facilitating the validation of an argument, and the evaluation of its strength, veracity and applicability.

Named Entity Recognition (NER)

Extraction of entity mentions over the arguments.

- A Greek BERT model [1] was fine-tuned for the NER task.
- elNER [2], a manually annotated Greek corpus was used as a corpus for the fine-tuning and the evaluation.
- state-of-the-art performance: 91% micro-average F1.
- trained model is open-source and available in `https://gitlab.isl.ics.forth.gr/papanton/debatelab-nel`

Candidate Generation

For each entity mention, we generate a list of possible entities from a multitude of wiki-based KBs.

- Several lookup queries to offline and online services.
- Alternative queries through lemmatizing, stemming and abbreviation expansion of Greek first names sent to lookup services.

1. **Wikipedia & DBPedia:** offline alternative names and redirect pages from the Greek Wikipedia.
2. **Wikidata:** queries through the wbsearchentities service.
3. **Wikifier:** retrieving Wikipedia additional candidates.
4. **Python Wikipedia library:** Wikipedia additional candidates employing the search mechanism of the Greek Wikipedia.

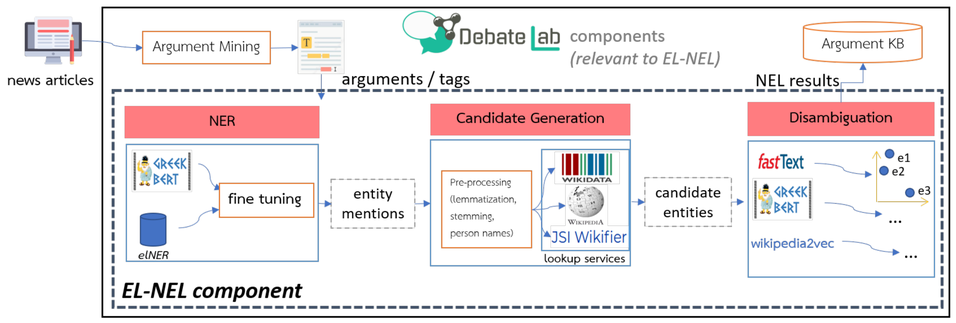

Figure 1:Figure caption

Disambiguation

- 3 different types of entity embeddings, two context-independent i.e., **fastText**, **wikipedia2vec** and one context-dependent that of the transformers-based **BERT**.
- For each embedding method, we compute the cosine similarities of the candidates in a k-partite graph, where k is the number of entity mentions, and candidates for the same entity mention belong to the same partition.

Extending to other languages

- The proposed pipeline is applicable to any language with at least a medium-resource coverage (i.e., NER corpora and/or pre-trained NER models, POS taggers, a relevant sizeable Wikipedia corpus)
- Wikidata since it uses the same entity identifiers across all available languages, could be used as a multilingual hub for entities.

Conclusion

- We present EL-NEL, a Greek entity linking component which is employed in the DebateLab project.
- Ongoing work includes:
  - exploration of Knowledge Graphs beyond Wiki-based ones.
  - incorporation of graph embeddings (e.g., ED-GNN) for better capturing the context of entity mentions
  - suitable approximation to the NP-hard problem of disambiguation multiple entities in the same argument.

References

[1] John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos.
GREEK-BERT: the greeks visiting sesame street.
In *SETN*, 2020.

[2] Nikos Bartziokas, Thanassis Mavropoulos, and Constantine Kotropoulos.
Datasets and performance metrics for greek named entity recognition.
In *SETN*, 2020.

Acknowledgements

This project has received funding from the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under grant agreement No 4195.

Contact Information

- Email: papanton@ics.forth.gr