EL-NEL: Entity Linking for Greek News Articles

Katerina Papantoniou[1,2], Vasilis Efthymiou[1], and Giorgos Flouris[1]

[1] Institute of Computer Science, FORTH, Greece
[2] Department of Computer Science, University of Crete, Greece
{papanton|vefthym|fgeo}@ics.forth.gr

Abstract. Readers, as well as journalists, are overwhelmed with the information available in online news articles, making it very difficult to verify and validate their content. An important tool to support readers in this task is that of named entity linking (NEL), i.e., semantically annotating entities mentioned in text with entities described in knowledge bases (KBs). In this poster, we introduce EL-NEL, the NEL component of DebateLab, a project with the more general objective of representing, mining and reasoning with online arguments. The models trained in EL-NEL are offered as open source.

Keywords: Entity Linking · Entity Disambiguation.

1 Introduction

The current news ecosystem is characterised by a rapid spread of news and a vast increase in the quantity of information available. This abundance of information makes it difficult for readers to identify good-quality journalistic resources, which are set apart by the use of credible and justified information and opinions, in the form of well-formulated arguments based on facts. The problem is aggravated by the fact that the provided information and arguments are often conflicting, and it is very difficult for a reader to evaluate their credibility without access to the relevant factual resources. This difficulty often leads to unwanted behaviours, such as the placement of blind and unjustified trust to specific opinions or sources, the generation of echo chambers, confirmation bias, and increased radicalisation and polarisation.

The DebateLab project[1] 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 wishes to be well-informed about topics or entities of interest (e.g., persons, locations, events).

[1] https://debatelab.ics.forth.gr/

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.

This work describes EL-NEL, the NEL component of DebateLab. Although NEL is quite popular in the literature, and various pre-trained models exist, the vast majority of those works are only applicable to the English language. Indeed, many multilingual or language-agnostic approaches are usually based on resources that are of poor quality [12], and biased towards the English language [5]. EL-NEL performs NEL for the Greek language, the language considered by the DebateLab project.

More specifically, this poster describes a pipeline for non-English NEL, which employs alternative state-of-the-art tools (e.g., BERT [6], wikipedia2vec [13], fastText [3]) for its components. The generated resources are made publicly available as open source[2]. The goal of this work is not to merely showcase the preliminary results for Greek NEL, but rather to present the process followed, and, hopefully, help researchers studying NEL for languages other than English.

2 Approach

Figure 1 presents the architecture of EL-NEL. The first step in the workflow of EL-NEL is to receive the arguments and tags from a news article, as provided by *Argument Mining*, one of the components of DebateLab. We treat arguments and tags differently, since arguments are typically whole sentences, while tags are typically keywords entered by the article author, or extracted automatically.

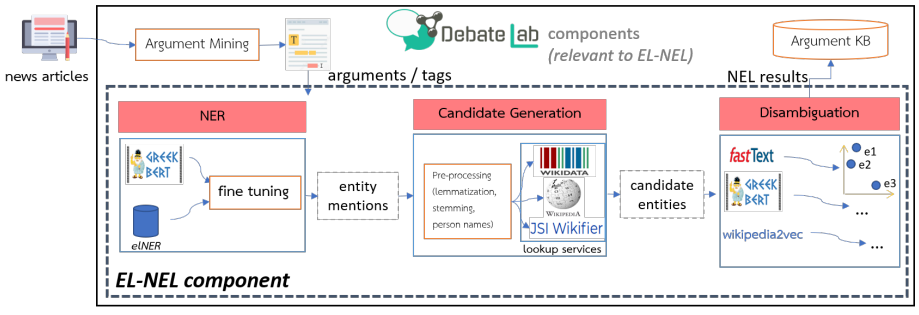

Fig. 1: The architecture of EL-NEL.

[2] https://gitlab.isl.ics.forth.gr/papanton/debatelab-nel

Named Entity Recognition (NER). In the case of arguments only (and not tags), we perform NER to extract entity mentions. We evaluated off-the-shelf NER tools (e.g., SpaCy[3], polyglot[4]), but we also fine-tuned a Greek BERT model [7]. For the fine-tuning, we used a manually annotated Greek corpus for NER (elNER [2]) consisting of 21,153 sentences from the news domain. The training was performed using the NERDA framework[5].

Our preliminary results, presented in Table 1, indicate that the fine-tuned BERT model (with a micro-averaged F1 score 91%) yields a negligible improvement over a SpaCy model (90% micro-averaged F1) that we trained on the same Greek corpus (elNER), and a significant improvement over the pre-trained Greek SpaCy model (51% micro-averaged F1).

Table 1: Evaluation results on elNER4 dataset. BERT stands out for our model based on fine-tuned Greek BERT, SpaCy is the model produced by training SpaCy on the elNER- datasets, and SpaCy pre is the pre-trained SpaCy model for Greek.

	BERT			SpaCy			SpaCy pre		
	P	R	F1	P	R	F1	P	R	F1
Location	0.93	0.93	0.93	0.90	0.92	0.91	0.75	0.01	0.01
Organization	0.89	0.89	0.89	0.90	0.85	0.87	0.78	0.57	0.66
Person	0.96	0.97	0.97	0.94	0.95	0.95	0.88	0.85	0.86
Miscellaneous	0.77	0.81	0.79	0.79	0.79	0.79	-	-	-
Micro average	0.91	0.92	0.91	0.90	0.89	0.90	0.52	0.44	0.48
Macro average	0.89	0.90	0.90	0.88	0.88	0.88	0.80	0.47	0.51

Candidate Generation. For each entity mention, we generate a list of possible entities from a multitude of wiki-based KBs (Wikidata, DBpedia, Wikipedia, YAGO). This component is mostly relying on existing lookup services offered by those KBs. For each entity mention, we perform several lookup queries to those services and rank the generated candidates based on the employed service and the pre-processing applied to each query (e.g., lemmatizing using TreeTagger [10] and Stanza [9], stemming using a snowball stemmer [8], and expanding abbreviated first names using a list of common Greek names).

Wikidata. We look up the entity mentions in Greek using the *wbsearchentities* service and retrieve the top-k results per query (k empirically set to 5). If we get no lookup results, we try again first with the lemmatized and then the stemmed version of the original query (i.e., entity mention). The order of executing those lookup queries is maintained in the output rankings of the candidate entities.

[3] `https://spacy.io/`
[4] `https://github.com/aboSamoor/polyglot`
[5] `https://github.com/ebanalyse/NERDA`

Wikipedia and DBpedia. We extract offline alternative names and redirect pages from the Greek Wikipedia, which we add to the list of candidate entities for each entity mention. In addition, the search mechanism of Wikipedia was also employed and the high ranked results were added to the candidate set. For DBpedia candidates, we simply use the Wikipedia suffixes.

Wikifier. We get additional Wikipedia candidates by calling the Wikifier [4] lookup service for each entity mention, and retrieving the top candidate. This tool enables multilingual semantic annotation of text with links to Wikipedia articles through a PageRank-based approach for disambiguation.

Disambiguation. We distinguish between two cases for entity disambiguation: (*i*) NER has detected a single entity mention, and we return the results obtained by a single ranked list of candidates, and (*ii*) more than one entity mentions are detected, so we consider the cohesion of the candidates, returning those whose entity embeddings are more similar to each other.

We examine three different types of entity embeddings, two context-independent (fastText [3], wikipedia2vec [13]) and one context-dependent (BERT [6]) based on transformers. 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.

In our preliminary experimental results, we observe that BERT performs best for the case of single entity mentions, since it is the only context-dependent method, while fastText performs better when more than one entity mentions appear in an argument, since in that case we can also consider the other entity mentions as context.

Extending this approach to other languages. The proposed pipeline can be extended to other languages too, as long as a medium-resource coverage (i.e., NER corpora and/or pre-trained NER models, POS taggers, a relevant sizeable Wikipedia corpus) is available in the target language. The adoption of the approach in low-resource languages is also applicable and can benefit from cross-lingual transfer learning approaches [1] or unsupervised learning approaches. However, in all these cases an extra provision must be taken for the idiosyncrasies of each language such as morphology, syntax etc. Finally, we note that, unlike other target KBs, Wikidata uses the same entity identifiers across all available languages.

3 Conclusions and Future work

In this poster, we introduced EL-NEL, a Greek named entity linking component which is employed in the DebateLab project. We have briefly described the modular architecture and alternative options for each component of EL-NEL, as well as some preliminary findings. In our ongoing and future work, we plan to explore KGs beyond the Wiki-based ones. We also plan to incorporate graph embeddings (e.g., ED-GNN [11]) for better capturing the context of entity mentions, and to find a suitable approximation to the NP-hard problem of disambiguating multiple entity mentions in the same argument.

Acknowledgement

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.

References

1. Alyafeai, Z., AlShaibani, M.S., Ahmad, I.: A survey on transfer learning in natural language processing. CoRR **abs/2007.04239** (2020)
2. Bartziokas, N., Mavropoulos, T., Kotropoulos, C.: Datasets and performance metrics for greek named entity recognition. In: SETN (2020)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguistics **5**, 135–146 (2017)
4. Brank, J., Leban, G., Grobelnik, M.: Annotating documents with relevant wikipedia concepts. In: SiKDD (2017)
5. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: ACL. pp. 8440–8451 (2020)
6. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (2019)
7. Koutsikakis, J., Chalkidis, I., Malakasiotis, P., Androutsopoulos, I.: GREEK-BERT: the greeks visiting sesame street. In: SETN (2020)
8. Ntais, G., Saroukos, S., Berki, E., Dalianis, H.: Development and enhancement of a stemmer for the greek language. In: PCI (2016)
9. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A Python natural language processing toolkit for many human languages. In: ACL (2020)
10. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: ICNMLP (1994)
11. Vretinaris, A., Lei, C., Efthymiou, V., Qin, X., Özcan, F.: Medical entity disambiguation using graph neural networks. In: SIGMOD (2021)
12. Wali, E., Chen, Y., Mahoney, C., Middleton, T., Babaeianjelodar, M., Njie, M., Matthews, J.N.: Is machine learning speaking my language? A critical look at the nlp-pipeline across 8 human languages. CoRR **abs/2007.05872** (2020)
13. Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., Matsumoto, Y.: Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. In: EMNLP (2020)