Automating Benchmark Generation for Named Entity Recognition and Entity Linking

Katerina Papantoniou, Vasilis Efthymiou, and Dimitris Plexousakis

FORTH-ICS, Greece
{papanton,vefthym,dp}@ics.forth.gr

Abstract. Named Entity Recognition (NER) and Linking (NEL) have seen great advances lately, especially with the development of language models pre-trained on large document corpora, typically written in the most popular languages (e.g., English). This makes NER and NEL tools for other languages, with fewer resources available, fall behind the latest advances in AI. In this work, we propose an automated benchmark data generation process for the tasks of NER and NEL, based on Wikipedia events. Although our process is applied and evaluated on Greek texts, the only requirement for its applicability to other languages is the availability of Wikipedia events pages in that language. The generated Greek datasets, comprising around 19k events and 41k entity mentions, as well as the code to generate such datasets, are publicly available.

1 Introduction

We are witnessing a proliferation of news articles available on the Web, making it difficult for readers to identify good-quality journalism with well-formulated and factually supported arguments. Despite the abundance of news articles available, it is still challenging to retrieve information related to a specific entity of interest (e.g., person, event, organization), in order to compare the arguments in favor of, or against a specific claim about them and shape an informed opinion. This often leads to easy spread of misinformation and conspiracy theories, sometimes with huge political, socio-economical or health impact. The DebateLab project[1] is conducting research towards representing, mining and reasoning with online arguments. The goal of this project is to offer a suite of tools and services that will assist both professional journalists in accomplishing everyday tasks, and readers who wish to be well-informed about topics or entities of interest.

A main component of DebateLab is EL-NEL [14], a tool responsible for *named entity linking (NEL)* in Greek news articles. NEL is the task of mapping parts of a text (called *entity mentions* or *surface mentions*) to uniquely identified entity descriptions provided in a target knowledge graph (KG). Consequently, it requires a step of detecting such entity mentions, along with their possible entity type, in the given texts, a task known as *named entity recognition* (NER).

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

Although NEL tools have recently seen great advances with language models pre-trained on large document corpora in English, few models exist for less popular languages. Worse yet, those few models are trained and evaluated on document corpora of much smaller scale, and tested by far fewer people than their popular-language counterparts. Due to those challenges, NEL tools in those low-resource languages fall far behind the latest advances in AI. To deter this gap, in this work, we propose an automated, language-agnostic benchmark data generation process for NER and NEL tasks using Wikipedia events pages[2]. We use the manually curated links in those event pages as the ground truth for NER and NEL tools and show how existing tools can be evaluated using such data. Although we use Wikipedia events in the Greek language as a use case, and release our data in Greek, our publicly available source code is language-agnostic and can be easily used for other languages.

It is not until recently that benchmark dataset for Greek NER have become publicly available (e.g., a NER dataset for spaCy [10], and a manually annotated corpus of Greek newswire articles, elNER [4]). Among the works that adopt multilingual approaches for the creation of benchmark datasets for NER based on Wikipedia articles, only two (Polyglot [2] and WikiAnn [13]) include Greek and only one (WikiAnn) also covers the NEL task. Unlike those benchmarks, the datasets generated by our method are news-/events-oriented. MEANTIME [11] and [8] are also targeting news/events. The MEANTIME [11] corpus consists of 120 manually annotated English Wikinews articles. [8] extracted 170k events from Wikipedia event pages from 9 languages (not including Greek). The main differences of our work compared to [8], are that we attribute entity types to the recognized entities, enabling its applicability to state-of-the-art NER tools, and that we perform a data enrichment step to fix red links.

In summary, the contributions of this work are the following:

- We offer publicly available benchmark datasets[3] for NER and NEL in Greek, with a permissive license (CC BY-SA 3.0), generated from Wikipedia events.
- We open-source the code[4] that generated those datasets, which can be adapted to generate similar datasets in more languages and, potentially, from more data sources.
- We show that these benchmarks can be used to evaluate existing NER and NEL tools, posing new challenges for such tools.

2 Methodology

In this section, we describe the methodology for the construction of the NER and NEL benchmark data.

Data Extraction. We extracted Greek Wikipedia events, redirects, mappings between Wikidata and Wikipedia articles, the Wikidata type(s) for each

[2] Greek wikinews page is not currently active. We expect that this may also be the case for many other languages.

[3] https://zenodo.org/record/7429037

[4] https://gitlab.isl.ics.forth.gr/debatelab/elwiki_events_benchmark

article, and the Greek Wiktionary, between 2009 (when the first Greek Wikipedia Events appeared in the template used today) and 2022.

Cleaning and Filtering. Our cleaning and filtering process includes the attribution of surface mentions to specific entity types, the removal of surface mentions not related to those types, and the concatenation of consecutive sentences referring to the same event. For the latter, we rely on heuristics. The entity types selected in this work are: event, facility, geopolitical entity, location, organization, person, product and work of art. For the attribution for each annotation to the entity types above, we use the instanceOf property of Wikidata. **Data Enrichment.** In addition to cleaning and filtering, we also enrich the extracted data, by filling in some of the so-called "red links", i.e., links to non-existing Wikipedia pages. Instead of disregarding red links, we tried to match such links with the corresponding (language-agnostic) Wikidata identifier, in order to increase the linkage of this dataset. For the collected dataset, approximately 14K out of 64K (∼21%) links were originally red links. We managed to recover ∼10K of those red links, by following the processing steps described below: (*i*) lexical transformations, e.g., convert first letter of placeholder suffix to uppercase, reorder words within the same entity mention, (*ii*) using the surface mention text as a wikidata search query, and (*iii*) translating the surface mention to English, using M2M100 [6], and repeating the same process in English.

3 Experiments

In this section, we evaluate the following NER methods: *EL-NEL* [14], *Neural-ILSP* [16], *NLP-AUEB* [17], *spaCy* [9], and *Polyglot* [2]; then, we evaluate the following NEL methods: *EL-NEL* [14], *WAT* [15], *spaCy fishing* [1], and *ReFinED* [3]. For all NEL tools, except *EL-NEL* which supports Greek, we translate the texts and the surface mentions in English to get the results.

NER Evaluation Methodology. For NER, we follow the so-called partial evaluation schema [5], that defines the following cases: A correct case (COR), when the surface mention and entity type in the ground truth match exactly with the NER tool. A partially correct case (PAR), when the surface mention in the ground truth overlaps with the surface mention returned by the NER tool (ignoring entity type). A missing case (MIS), when a ground truth annotation is not returned at all by a NER tool. A spurious case (SPU), when a NER tool suggest an annotation that does not exist in the ground truth. Then, precision and recall are defined as $Precision = \frac{COR+0.5\ PAR}{COR+PAR+SPU}$, and $Recall = \frac{COR+0.5\ PAR}{COR+PAR+MIS}$.

NEL Evaluation Methodology. For NEL, we follow the spaCy scorer[5] approach that considers only the links provided for entity mentions that overlap with the NER ground truth. We consider the following cases: A true positive (TP) occurs when a NEL tool suggests the same link as the ground truth. A false positive (FP) occurs when a NEL tool suggests a different link than the correct link. This includes the case of a NEL tool suggesting any link for a

[5] https://github.com/explosion/spaCy/blob/master/spacy/scorer.py

known red link in the ground truth. A false negative (FN) occurs when a NEL tool returns no link (or an incorrect link) for an entity mention that appears in the NEL ground truth (not a red link). True negatives (TN) are ignored. As a consequence, precision and recall are almost identical, so we report only the micro- and macro-averaged F1-scores. Due to the imbalance of the entity types, F1-micro that takes into account proportion of every type is more meaningful.

NER Evaluation Results. The results of NER, presented in Table 1, show that the NER benchmark is challenging for all evaluated methods, with the highest F1 being 0.56 for the EL-NEL system. Polyglot, which detects a small portion of the entity types (which are, nonetheless, the most representative in the dataset), shows the highest precision, but also the lowest recall. The low recall for all the tools is explainable due to spurious cases, since there is a gap between what the Wikipedians choose to annotate and the tools' predictions, which are usually more exhaustive.

Table 1. NER results.

Tool	Types	Precision	Recall	F1
EL-NEL	8	0.76	0.46	0.56
Neural-ILSP	8	0.51	0.46	0.46
NLP-AUEB	8	0.84	0.34	0.47
spaCy	6	0.75	0.43	0.53
Polyglot	3	0.86	0.33	0.47

Table 2. NEL results.

Tool	F1-micro	F1-macro
EL-NEL	0.91	0.82
WAT	0.96	0.96
SpaCy fishing	0.77	0.78
ReFinED	0.95	0.90

NEL Evaluation Results. The results of NEL are shown in Table 2. Overall, WAT shows the best performance among all NEL tools, followed by the neural-based ReFinED. A more detailed analysis, as presented in Table 3, reveals that all NEL tools struggle with events, while they show impressive results for persons. WAT does not categorize surface mentions into entity types, so it is skipped from this table.

Table 3. NEL results (F1-macro) per entity type.

	EL-NEL	SpaCy fishing	ReFinED
EVENT	0.61	0.47	0.78
FAC	0.78	1	0.88
GPE	0.92	0.72	0.96
LOC	0.92	0	0.94
ORG	0.9	0.81	0.94
PERSON	0.92	0.92	0.94
PRODUCT	0.58	1	1
WORK_OF_ART	1	0.6	0.71

In terms of computational time, SpaCy fishing and ReFinED that do not require calls to web APIs are significantly faster (few minutes vs a few hours). An efficiency evaluation of the tools falls beyond the scope of this work.

4 Conclusion and Future Work

In this work, we have presented an open-source benchmark dataset for NER and NEL, built from Greek Wikipedia Events. The dataset consists of 24k sentences

and includes 41k entity mentions. We plan to follow the same methodology and release similar benchmark datasets in more languages. We also plan an experimental comparison with other benchmark datasets and models, as well as the enrichment of the collection with annotations that Wikipedians have not considered in their annotations. For the latter the structure of Wikipedia (e.g., anchor links, disambiguation pages) as well as co-reference information can be exploited as in [12] and [7].

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. entity-fishing. https://github.com/kermitt2/entity-fishing
2. Al-Rfou, R., Kulkarni, V., Perozzi, B., Skiena, S.: POLYGLOT-NER: massive multilingual named entity recognition. In: SIAM. pp. 586–594 (2015)
3. Ayoola, T., Tyagi, S., Fisher, J., et al.: ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In: NAACL. pp. 209–220 (2022)
4. Bartziokas, N., Mavropoulos, T., Kotropoulos, C.: Datasets and performance metrics for greek named entity recognition. In: SETN. pp. 160–167 (2020)
5. Chinchor, N., Sundheim, B.: MUC-5 evaluation metrics. In: MUC. pp. 69–78 (1993)
6. Fan, A., Bhosale, S., Schwenk, H., et al.: Beyond english-centric multilingual machine translation. J. Mach. Learn. Res. **22**, 107:1–107:48 (2021)
7. Ghaddar, A., Langlais, P.: Winer: A wikipedia annotated corpus for named entity recognition. In: IJCNLP. pp. 413–422 (2017)
8. Hienert, D., Wegener, D., Paulheim, H.: Automatic classification and relationship extraction for multi-lingual and multi-granular events from wikipedia. In: DeRiVE. pp. 1–10 (2012)
9. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017)
10. Ioannis Daras, Markos Gogoulos, P.L.: gsoc2018-spacy. GitHub (2018), https://github.com/eellak/gsoc2018-spacy
11. Minard, A., Speranza, M., Urizar, R., et al.: MEANTIME, the newsreader multilingual event and time corpus. In: LREC (2016)
12. Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R.: Learning multilingual named entity recognition from wikipedia. Artif. Intell. **194**, 151–175 (2013)
13. Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., Ji, H.: Cross-lingual name tagging and linking for 282 languages. In: ACL. pp. 1946–1958 (2017)
14. Papantoniou, K., Efthymiou, V., Flouris, G.: EL-NEL: Entity linking for greek news articles. In: ISWC Posters, Demos and Industry Tracks (2021)
15. Piccinno, F., Ferragina, P.: From tagme to WAT: a new entity annotator. In: ERD@SIGIR. pp. 55–62 (2014)
16. Prokopidis, P., Piperidis, S.: A neural NLP toolkit for greek. In: SETN. pp. 125–128 (2020)
17. Smyrnioudis, N.: A Transformer-based natural language processing toolkit for Greek–Named entity recognition and multi-task learning. BSc thesis, AUEB (2021)