EAL: A Toolkit and Dataset for Entity-Aspect Linking

Federico Nanni, Jingyi Zhang, Ferdinand Betz, Kiril Gashteovski
Data and Web Science Group, University of Mannheim - Germany
federico,kiril@informatik.uni-mannheim.de
jizhang,fbetz@mail.uni-mannheim.de

ABSTRACT

We present a toolkit and dataset for entity-aspect linking. The tool takes as input a sentence and provides the most relevant aspect for each mentioned entity; it is implemented in Python and available as a script and via an online demo. It is accompanied by the first large dataset of entity-aspects, comprising more than 20,000 entities manually linked to the most relevant aspect, given a sentence as context. Each is expressed in structured manner as Open Information Extraction (OIE) triples *(Subject, Relation, Object)*, having semantic information for polarity, modality, quantity and attributions.

KEYWORDS

entities, entity-aspects, open information extraction, information retrieval, knowledge base

ACM Reference Format:
Federico Nanni, Jingyi Zhang, Ferdinand Betz, Kiril Gashteovski. 2019. EAL: A Toolkit and Dataset for Entity-Aspect Linking. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'19).* ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

The availability of entity linking technologies provides a novel way to organize, categorize, and analyze large textual collections in digital libraries, by detecting mentions of entities in context and linking them to an entry in a knowledge base (e.g. Wikipedia, DBpedia, YAGO). However, in many situations a link to an entity offers only relatively coarse-grained semantic information, for instance when the entity is related to several different events, topics, roles, and – more generally – when it has different aspects. To address this issue, in our previous work [4] we have introduced the task of entity-aspect linking: given a mention of an entity in a contextual passage, we refine the entity link pointing to its Wikipedia page with respect to the aspect (i.e., the section) of the entity it refers to. For instance, the sentence:

> PM **Theresa May** has refused to say how she would vote if Britain held another EU referendum.

The entity Theresa May would be linked to the section/aspect Political_positions present on her Wikipedia page,[1] which specifically focuses on Brexit.

In our research, we have shown that using entity-aspect linking leads to significant and consistent improvements on different entity prediction and categorization tasks relevant for the digital library community [4, 5].

This paper reports on three contributions for supporting further research on entity-aspect linking: *a)* we release a Python implementation of our entity-aspect linking approach (EAL) and an easy-to-use online interface for performing entity-aspect linking in a single step *b)* we present the first large dataset of aspect links (EAL-D), which comprises over 20,000 entities that have been manually linked to its most related aspect by Wikipedia curators. This drastically extends our previous benchmark of 200 aspect links; *c)* on this new collection, we re-evaluate our entity-linking strategy previously presented, obtaining consistent results.

2 TOOLKIT

EAL is implemented in Python.[2] It takes an entity in context as input and returns the most relevant aspect, relying on the following main functions:

(1) **wiki-crawler**. This takes as input a Wikipedia entity and returns a aspect-dictionary, with section-headings as keys and section-contents as values. To do so, it currently crawls such information directly from Wikipedia, but we will extend it soon to process a Wikipedia dump to avoid the crawling bottle-neck. Following practices from the TREC-CAR organizers, we remove sections with titles such as "See Also", "External Links", etc.[3] and aggregate the initial paragraphs under the name "lead paras".
(2) **ranking.** This takes an aspect-dictionary and textual content (i.e., a sentence) as inputs and returns the most relevant aspect. Currently, it does so measuring the tf-idf content similarity between the sentence and the aspect; this was the single best performing feature presented in our previous work (see Table 3 in [4]), which gave us consistent results on the new large-scale dataset (see Table 1). We additionally plan to offer the possibility of choosing the other ranking functions presented in our previous work. In case no sections are present on the entity page, we return the "lead paras" as a general overview.

The entire pipeline is available through an online demo (see Figure 1): this takes a sentence as input, conducts entity linking using TagMe [1], performs EAL on all linked entities and finally provides back the most relevant aspect for each of them.[4] When the input context is longer than a sentence, as a first step our pipeline performs sentence tokenization with NLTK and then processes one sentence at a time. That is because we have provided evidence in our previous work that EAL performance drops with larger and

[1] https://en.wikipedia.org/wiki/Theresa_May#Political_positions

[2] The tool is available here: https://github.com/jinggz/Master-Thesis-EAL/tree/service
[3] See Phase 4: http://trec-car.cs.unh.edu/process/dataselection.html
[4] The online demo is available here: http://tools.dws.informatik.uni-mannheim.de/eal

JCDL'19, June 2019, Urbana-Champaign, Illinois USA

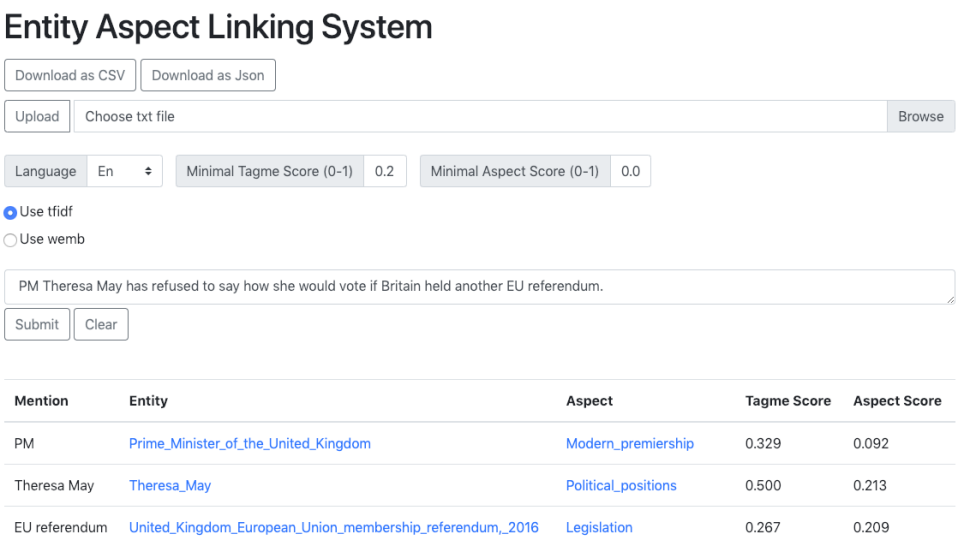

Figure 1: Snapshot of the demo interface: given a sentence it provides entity and aspect linking annotations.

noisier contexts, compared to a single sentence. EAL results are displayed and available for download as .json or .csv files. Relying on TagMe, we currently offer entity linking in English, Italian and German. We will offer aspect linking in these three languages in near future as well.

3 DATASET (EAL-D)

As discussed in our previous work, we rely on the fact that in Wikipedia's Manual of Style and Linking, contributors are encouraged to point hyperlinks to the specific section/aspect addressed in text when present on the entity page. In order to leverage this data, we use OPIEC[5] [3] – the largest OIE corpus to date, containing more than 341M OIE triples – which was created by running MinIE [2] on the entire English Wikipedia. OPIEC retains the golden entity and entity-aspect links from within the Wikipedia articles (which have been created by Wikipedia contributors). We use a subset of OPIEC, containing 3K triples with aspect links on both the subject and the object and 20K/29K triples with aspect links on at least the subject/object.

As done in our previous work, we remove sentence duplicates and we further exclude entity links pointing to redirecting pages and lists, entity and entity aspects which are not available anymore in Wikipedia and pages having only one aspect. For aspect-links pointing to a sub-section (h3 heading), we replace this with the main section (h2). The final dataset, consisting of around 8,000 sentences having an entity-aspect link as a subject of the extracted relation, 13,000 sentences having it as an object and 1,000 as both, is shared with the community to foster futher work on the topic.[6]

4 EVALUATION

To re-evaluate the performance of EAL, we present the results of our pipeline on our new large dataset (EAL-D) in comparison with the ones obtained in our previous work on a smaller, manually curated resource of 200 aspect links. The experiment studies the quality of EAL on linking the correct aspect of an entity, when only a sentence is given as a context and the entity is central to the sentence (i.e., it is the subject of the relation). The results presented in Table 1 reveal consistent performance of our strategy on the new dataset in comparison with our previous work.

Table 1: Precision at 1 (P@1) for aspect-linking using sentence context: comparison between performance on original and extended dataset.

Model	P@1 on [4]	P@1 on EAL-D (subj)
Header: tf-idf (cs)	0.44	0.46
Header: BM25	0.42	0.45
Content: tf-idf (cs)	0.62	0.60
Content: BM25	0.60	0.64

5 CONCLUSIONS

In this paper we presented a toolkit, online demo and large dataset to foster the adoption and improvement of entity-aspect linking technologies in the research community.

REFERENCES

[1] Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proc. of CIKM*.
[2] Kiril Gashteovski, Rainer Gemulla, and Luciano Del Corro. 2017. MinIE: Minimizing Facts in Open Information Extraction. In *Proc. of EMNLP*.
[3] Kiril Gashteovski, Sebastian Wanner, Sven Hertling, Samuel Broscheit, and Rainer Gemulla. 2019. OPIEC: An Open Information Extraction Corpus. In *Proc. of AKBC*.
[4] Federico Nanni, Simone Paolo Ponzetto, and Laura Dietz. 2018. Entity-Aspect Linking: Providing Fine-Grained Semantics of Entities in Context. In *Proc. of JCDL*.
[5] Federico Nanni, Simone Paolo Ponzetto, and Laura Dietz. 2018. Toward comprehensive event collections. *International Journal on Digital Libraries* (2018).

[5] https://www.uni-mannheim.de/dws/research/resources/opiec/
[6] https://federiconanni.com/eal-d/