Standardization challenges to a better evaluation in entity normalization

Arnaud Ferré
Université Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France

1 A non-standardized definition

The overall goal of entity normalization is to link identified entity mentions to standard entities from an available set of unambiguous references (ontology, terminology, thesaurus, dictionary, …). The entity mentions are possibly represented by multi-word non-contiguous expressions. This task commonly assumes an entity recognition was firstly made. The normalization task then consists in linking these identified mentions of interest to zero, one or several standard entities. We propose this relatively general definition, whereas in practice, more constrained ones are used.

The task itself behind "*entity normalization*" can be also named concept normalization or entity linking/disambiguation or even entity/concept grounding. Moreover, there may indeed be some subtle variations in their definition (Martinez-Rodriguez et al., 2020). For instance, some consider that "entity linking" refers to the overall task of entity recognition and entity disambiguation (Kolitsas et al., 2018), while others consider that it is similar to entity disambiguation only (Derczynski et al., 2015). Moreover, some consider indeed entity linking and entity normalization as synonyms (Chen et al., 2021). It seems that the difference stems from the emergence of this issue in different NLP communities, whose different contexts have led to differences in the difficulty of approaching the task.

2 A non-standard scoring metric

The consensus evaluation metric used at the task level is the "*accuracy*", which is basically the average of a strict metric over all evaluated mentions. But if there is no online evaluation platform or independent evaluation programs, which is mainly the case, authors compute the scores for their methods by themselves. As there are some subtle variations between datasets (e.g. multi-entities normalization), it is very likely that everyone does not use the exact same scoring function, which we show that it can imply different scores for the same method on the same dataset.

3 A non-standard and biased evaluation

Manually annotated corpora with standard entities are created for evaluating normalization methods. Annotations by domain experts identify mention boundaries and associate concepts from a chosen set of standard entities. The annotated corpus is split into at least a training set for method optimization and a test set for performance estimation. However, a blind spot is the study of overlaps between these sets, which can sometimes lead to a majority of examples being present in the test set already encountered in the train/dev sets. In the same way, the distribution of mentions among the classes/standard entities can be unrealistic in order to limit cases of few- or zero-shot learning.

Other important biases persist, such as the fact that not everyone uses the same annotation reference. In particular, it is possible to use for classification only the standard entities appearing in the test set, rather than all the entities addressed by the task. As a result, on the same dataset, we show that the same method can artificially obtain a higher score by decreasing the number of standard entities.

4 Non-standard practice

Even by agreeing on a single evaluation measure, and in a well-defined context, we were able to show that a simple number for accuracy was not enough to give a real idea of a method's performance. Indeed, contemporary neural methods require a random initialization of its parameters, and depending on this initialization, we can already observe more or less significant variations in the results. However, in many cases, no information on method variability is provided.

5 Acknowledgments

This abstract is mainly based on the work from:

Arnaud Ferré and Philippe Langlais. 2023. An analysis of entity normalization evaluation biases in biomedical domains: What is wrong? What can we do better? *BMC bioinformatics*, vol. 24, no 1, p. 227.

6 References

Lihu Chen, Gaël Varoquaux, and Fabian M Suchanek. 2021. A lightweight neural model for biomedical entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12657–12665.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of Named Entity Recognition and Linking for Tweets. *Information Processing & Management*, 51(2):32–49. arXiv: 1410.7182.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-End Neural Entity Linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529.

Jose L Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. 2020. Information extraction meets the semantic web: a survey. *Semantic Web*, 11(2):255–335.