What should Entity Linking link?

Henry Rosales-Méndez, Barbara Poblete and Aidan Hogan

Millenium Institute for Foundational Research on Data
Department of Computer Science, University of Chile
{hrosales,bpoblete,ahogan}@dcc.uchile.cl

Abstract. Some decades have passed since the concept of "named entity" was used for the first time. Since then, new lines of research have emerged in this environment, such as linking the (named) entity mentions in a text collection with their corresponding knowledge-base entries. However, this introduces problems with respect to a consensus on the definition of the concept of "entity" in the literature. This paper aims to highlight the importance of formalizing the concept of "entity" and the benefits it would bring to the Entity Linking community, in particular relating to the construction of gold standards for evaluation purposes.

1 Introduction

Entity Linking (EL) is a task in Information Extraction that links the entity mentions in a text collection with their corresponding knowledge-base (KB) entries. With EL, we can take advantage of a large amount of information available in publicly available KBs (e.g., Wikipedia, DBpedia, Wikidata) about real-world entities and their relationships to obtain semantic information that can be used to achieve a better understanding of text corpora. For example, in the text "*Michael Jackson was born in Gary, Indiana*", if we can link the mention *Michael Jackson* with its entry in Wikidata (`http://www.wikidata.org/entity/Q2831`), then we know the text is about a U.S. pop singer and can provide further KB facts about that singer to the reader, etc. Along these lines, EL has a wide range of applications, including semantic search, semantic annotation, text enrichment, relationship extraction, entity summarization, etc.

The EL task can be broken down into two main sub-tasks. First, entity mentions must be located in the text (referred to as "recognition"). Second, those mentions must be associated with a suitable identifier from the KB (referred to as "disambiguation"). The overall process can be complicated by a number of factors. One obstacle, for example, is name variations, where, e.g., *Michael Jackson* can be referred to by his full name *Michael Joseph Jackson*, or also by *Michael* or *Jackson* or *M. Jackson*, etc. Another major obstacle is ambiguity, where *Michael Jackson* can refer to a variety of musicians, actors, politicians, soldiers and scientists, but only one is the appropriate person. A more thorough review of EL systems can be found in the survey of Martinez et al. [1]

While the previous challenges for EL are well-known, another more fundamental issue is often overlooked by the community: the question of *what is an*

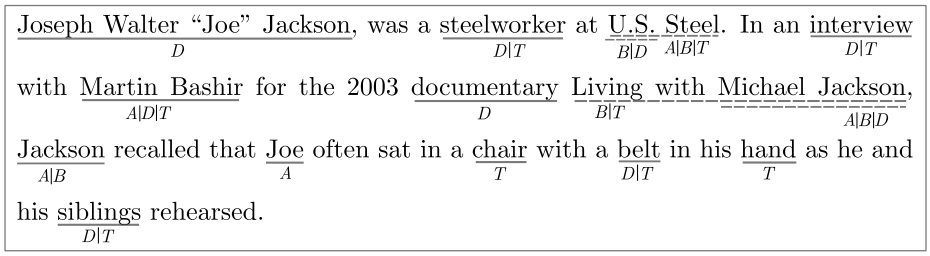

Fig. 1. Example annotations produced by four EL systems: AIDA (A), Babelfy (B), DBpedia Spotlight (D) and TAGME (T).

"entity"? Though several definitions have emerged about what an entity should be [2,3,4,5], there is, as of yet, no clear consensus [6,7].

This question has a major impact on EL research, leaving unclear which entity mentions in a text should be linked by EL systems or annotated by gold standards for evaluation purposes. To illustrate, Figure 1 shows an example text snippet from Wikipedia and the annotations produced by popular EL approaches: AIDA [8], Babelfy [9], DBpedia Spotlight [10] and Tagme [11]. Here we can see how these systems differ in their recognition of entities. Although most systems correctly recognize and link popular entity mentions like *Michael Jackson*, for no entity mention do all systems agree. The fundamental question then is: *which annotations are "correct"?* The answer depends on how "entity" is defined.

2 What is an "Entity"?

For the 6$^{\text{th}}$ Message Understanding Conference [2] (MUC-6), the concept of "*named entity*" was defined as those terms that refer to instances of proper-name classes such as *person*, *location* and *organization*, and also, to numerical classes such as *temporal expressions* and *quantities*. Many *named entity recognition* (NER) tools and training datasets/gold standards were developed to recognize and type entity mentions corresponding to these classes. However, researchers later became interested in Entity Linking (EL), where mentions were no longer simply recognized, but also linked to a reference KB (often using Wikipedia). Such KBs contain entities that do not correspond to traditional MUC-6 types so this definition was no longer exhaustive: in Figure 1, while the people and organizations would be covered under the MUC-6 consensus, the documentary "*Living with Michael Jackson*" would not; on the other hand, no system annotates "*2003*" from the MUC-6 class *Timex*.

Some authors have since defended the class-based proposal of MUC-6, incorporating new classes into the initial definition such as *products*, *financial entities* [12], *films*, *scientists* [13], etc. On the other hand, Fleischman [14] proposed to separate the classes into multiple specific subclasses (e.g., deriving *city*, *state*,

country from the class *location*). Different processes and models can then be applied for different entity types. In general, however, such class-based definitions are inflexible, where at the time of writing, a KB such as Wikidata has entities from 50,000 unique classes, with more classes being added by users. Hence some authors have preferred more general definitions, but these often lack formality [3,4]. For example, Ling et al. [7] use the definition "*substrings corresponding to world entities*", but this is cyclical: by using "*entity*" in the definition, it omits what should be considered an "*entity*" in the first place.

Another point of view is to define an entity based on what is described by a knowledge-base; e.g., Perera et al. [5] define an entity as those described by Wikipedia pages with no ambiguity. While this avoids class-based restrictions and offers a practical, operational definition for EL purposes, it too has issues. Entities are tied to a particular version of a KB, making it impossible to create general gold standards or to reflect *emerging entities* that may be added to the KB in future. Furthermore, Wikipedia has articles for general terms such as *documentary* and *belt*, though as per Figure 1, many tools and authors would not consider such terms as "entities", but rather as being general words/concepts (and thus the subject of a different task: Word Sense Disambiguation (WSD)).

Even if we establish a clear definition for "*entity*", we are still left to clarify what kinds of *entity mentions* should be recognised by EL. For example, all prior definitions agree that the singer *Michael Jackson* is an entity, but in the text of Figure 1, no definition clarifies whether or not an EL system should recognize and link the mentions *Jackson* (a *short mention*) and/or *he* (*a pronoun*) to the KB entity for *Michael Jackson* to which they refer; some authors, such as Jha et al. [15], consider this a task independent of EL called *Coreference Resolution* (CR), while others consider it part of EL to disambiguate entity types [16]. Furthermore, in the mention "*Living with Michael Jackson*", some authors consider the inner *overlapping mention* of "*Michael Jackson*" as valid [9,17]; others, such as Jha et al. [15], only consider the larger mention as valid.

In the context of EL, we thus see a lack of consensus, not only on the notion of an entity, but also on the notion of an entity mention; such disagreement may explain some of the differences in results for the four EL systems over the example text of Figure 1. But this lack of consensus undermines the possibility of collaborative research; for example, as suggested by Figure 1, datasets labeled under one definition should not, rightfully speaking, be used to train or evaluate tools developed under a different assumption; this, however, has been the case [17,15]. In different labeled datasets used in EL for training or evaluation purposes, we find that most datasets do not label overlapping entities nor coreferences; however, SemEval 2015 Task 13 does consider overlapping entity mentions [18], while the OKE Challenge 2016 and MEANTIME datasets annotate coreferences. In benchmarks, such datasets are then biased toward approaches adopting similar definitions for entities and mentions; e.g., one dataset may implictly mark overlapping entities (as produced by Babelfy in Figure 1) as true positives while others may mark them as false positives.

3 Proposed Solution

We are not the first to identify such issues: Ling et al. [7] draw similar examples on the lack of consensus on EL, while Jha et al. [15] also identify this problem and propose a set of rules to serve as best practices for benchmark creation. While standardizing the creation of EL benchmarks and making explicit the assumptions under which they are generated is a step in the right direction, as previously discussed, it is not clear what assumptions should, in reality, be adopted. Jha et al. [15] propose, for example, that overlapping mentions be omitted (and, in fact, refer to their inclusion as "*errors*") but as discussed, other authors (including Ling et al. [7]) disagree on this specific issue.

Our position is that the more fundamental question needing to be resolved in the context of EL is not the semantic question of "*what is an 'entity'?*", but rather the practical question of "*what should Entity Linking link?*". The answer to this latter question, we argue, depends heavily on the application. For the purposes of semantic search – for example, finding all documents about US singers – coreference is not so important since one mention of *Michael Jackson* in a document may be enough to establish relevance. On the other hand, for extracting relations between entities, many such relations may be expressed in text with pronouns. Likewise an EL process may choose to recognise and link mentions of terms such as "*singer*" to the KB to help to apply a more accurate (collective) disambiguation of neighbouring mentions such as "*Michael Jackson*" (as proposed by Babelfy). Any single set of rules or definitions by which EL should be conducted is, we thus argue, exclusionary and an oversimplification.

Hence our proposed solution is not to provide another unilateral definition of what EL should consider as an "*entity*" or an "*entity mention*", but rather to be explicit on the different forms of entities and entity mentions that a particular EL system may wish to recognize and link. This would involve creating labeled texts – for training and benchmarking – that make explicit the different forms of entity mentions present, be they proper names, other terms present in the KB, overlapping entities, or coreferences. Tools and evaluators may then choose to explicitly include/exclude whichever entity (mentions) they consider relevant for their application. Much like the original MUC-6 definitions, we propose that such labels should be established through consensus in the community and included in standards such as NLP Interchange Format (NIF) [19]. While this would add some additional complexity to the generation of labeled datasets and the processes of evaluation (when compared with, e.g., the proposals of Jha et al. [15]), we argue that such additional effort is no more than what the EL community will *require* as it matures. We would thus like to propose a metric that takes into account the ambiguity of what is an entity, and that measures the capacity of an EL system to link different types of entities.

Acknowledgements The work of Henry Rosales-Méndez was supported by CONICYT-PCHA/Doctorado Nacional/2016-21160017. The work was also supported by the Millennium Nucleus Center for Semantic Web Research under Grant NC120004.

References

1. Martinez-Rodriguez, J., Hogan, A., Lopez-Arevalo, I. Information Extraction meets the Semantic Web: A Survey. Semantic Web journal. 2018 (to appear)
2. Grishman, R., and Sundheim, B. Message understanding conference-6: A brief history. In COLING **1** (1996)
3. Eckhardt, A., Hreško, J., Procházka, J., Smrí, O. Entity linking based on the co-occurrence graph and entity probability. ERD, ACM (2014) 37–44
4. Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Ciravegna, F. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Journal of Web Semantics, **4**(1) (2006) 14–28
5. Perera, S., Mendes, P. N., Alex, A., Sheth, A. P., Thirunarayan, K. Implicit entity linking in tweets. In ISWC, (2016) 118–132
6. Borrega, O., Taulé, M., Martí, M.A. What do we mean when we speak about Named Entities. In Proceedings of Corpus Linguistics, 2007
7. Ling, X., Singh, S., Weld, D. S. Design challenges for entity linking. TACL **3** (2015) 315–328
8. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: EMNLP, ACL (2011) 782–792
9. Moro, A., Raganato, A., Navigli, R. Entity linking meets word sense disambiguation: a unified approach. TACL **2** (2014) 231–244
10. Mendes, P. N., Jakob, M., García-Silva, A., Bizer, C. DBpedia spotlight: shedding light on the web of documents. In I-Semantics (2011) 1–8
11. Ferragina, P., Scaiella, U.: Tagme: on-the-fly annotation of short text fragments (by Wikipedia entities). In: CIKM, ACM (2010) 1625–1628
12. Minard, A. L., Speranza, M., Urizar, R., Altuna, B., van Erp, M. G. J., Schoen, A. M., van Son, C. M. MEANTIME, the NewsReader multilingual event and time corpus. LREC-ELRA (2016)
13. Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., Daniel S. W., Yates, A. Unsupervised named-entity extraction from the web: An experimental study. Artificial intelligence, **165**(1) (2005) 91–134
14. Fleischman, M. Automated subcategorization of named entities. In ACL (2001) 25–30
15. Jha, K., Röder, M., Ngomo, A. C. N. All that glitters is not gold–rule-based curation of reference datasets for named entity recognition and entity linking. In ESWC (2017) 305–320
16. Durrett, G., Klein, D. A joint model for entity analysis: Coreference, typing, and linking. TACL **2** (2014) 477–490
17. Luo, G., Huang, X., Lin, C. Y., Nie, Z. Joint entity recognition and disambiguation. In EMNLP (2015) 879–888
18. Rosales-Méndez, H., Poblete, B., Hogan, A. Multilingual Entity Linking: Comparing English and Spanish. In LD4IE (2017)
19. Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. Integrating NLP using linked data. In ISWC, (2013) 98–113