Hunger for Contextual Knowledge and a Road Map to Intelligent Entity Linking

Filip Ilievski, Piek Vossen, and Marieke van Erp

Vrije Universiteit Amsterdam, The Netherlands
{f.ilievski,piek.vossen,marieke.van.erp}@vu.nl

Abstract. The task of entity linking (EL) is often perceived as an algorithmic problem, where the novelty of systems lies in the decision making process, while the knowledge is relatively fixed. As a consequence, we lack an understanding about the importance and the relevance of diverse knowledge types in EL. However, knowledge and relevance are crucial: following the Gricean maxim, an author relies on assumptions about the knowledge of the reader and uses the most efficient and scarce, yet understandable, level of detail when conveying a message. In this paper, we seek to understand the EL task from a knowledge and relevance perspective. We define four categories of contextual knowledge relevant for EL and observe that two of these are systematically absent in existing entity linkers. Consequently, many contextual cases, in particular long-tail entities, can never be interpreted by existing systems. Finally, we present our ideas on developing knowledge-intensive systems and long-tail datasets.

Keywords: Entity Linking, Context, Long Tail, Knowledge, Reasoning

1 Introduction

The task of Entity Linking (EL) anchors recognized entity mentions in text to their semantic representation, thus establishing identity and facilitating the exploitation of background knowledge, easy integration, and comparison and reuse of systems. Various EL approaches have been introduced in recent years [15, 2, 10, 17, 12, 1]. These systems optimize the semantic coherence of entities using probabilistic disambiguation. This often revolves around graph optimization over potential entity interpretations within one document with the goal of finding a minimal well-connected graph that contains at most one entity interpretation per mention. Alternatively, machine learning algorithms combine local and global features to score the fit of mention interpretations against mention training data based on popularity, string similarity and the level of association between two entities (cf. [12]). Both graph-based and machine learning methods use common knowledge bases (e.g. Wikidata), and, despite the non-ideal coverage and bias of these sources, the systems yield F_1-scores in the order of 60-70%.[1]

Recently, we demonstrated that these scores are largely due to the dominance of a limited number of popular entities [3, 6]. The accuracy of most probabilistic

[1] http://gerbil.aksw.org/gerbil/overview

algorithms is mainly based on test cases for which there is sufficient training data and background knowledge. We refer to these frequently mentioned entities as the *linguistic head*. Besides being frequent news topics, the mentions of head entities are also frequent and the mention-to-entity dominance is very high.

However, at the same time there is a vast amount of *long-tail entities*, each different and with low frequency, that usually remain hard to resolve for any contemporary system. Support for this claim can be found in the related task of Word Sense Disambiguation. Here, the system accuracy on the most frequent word interpretations is close to human performance, while the least frequent words can be disambiguated correctly in at most 1 out of 5 cases [13]. It is our conviction that the linguistic long tail can never be fully tackled with further algorithmic inventions because these long-tail interpretations appear only incidentally and it is unlikely there will ever be sufficient training data. Even if we would increase the training data, it is impossible to guess the a-priori distribution that applies to any actual test set across all the options [13].

Additionally, probabilistic approaches do not employ any mechanisms to exclude anomalous interpretations. This leads to an explosion of potential interpretations which are dominated by the most popular ones even though they often do not make any sense. In the NewsReader project, for example, the most popular detected entity in 2.3 million news articles from 2003-2015 about the car industry was *Abraham Lincoln*, demonstrating how dominance leads to wrong and impossible interpretations [16]. This problem becomes even more substantial when we switch from the popular world represented in Wikipedia to resolving long-tail entities, where the surface form ambiguity becomes too big to handle.[2] For instance, while *Ronaldo* can refer to only a few popular entities according to Wikipedia, the number of people in the world that (have) share(d) this name is many orders of magnitude greater. For current systems, it is extremely hard to deal with this reality, while humans have no problem understanding news mentioning some non-famous *Ronaldo*. As these long-tail interpretations are only relevant within a specific context (time, location, topic, community), we need contextual knowledge and reasoning in order to decide which make sense.[3]

In this position paper, we argue that the extreme ambiguity representing the long tail can only be addressed by robust reasoning over well-targeted, but rich, contextual knowledge. The road to intelligent EL thus requires a revision and extension of contextual knowledge, as well as an approach to dynamically acquire such knowledge for each long-tail case. We summarize the knowledge used by humans when reading, and adapt an existing knowledge classification [9] for a knowledge-intensive EL framework. We compare contemporary systems against this framework to assess which knowledge aspects are currently left out. We argue that our community should switch focus, away from systems and evaluations that

[2] Probabilistic methods are sensitive to even small changes in the background knowledge: only switching to a more recent Wikipedia version causes a drop in performance because of the increased ambiguity and the change in knowledge distribution [12].

[3] These differ from the domain-specific entities, which are defined through a single contextual dimension (of topic) and do not necessarily suffer from knowledge scarcity.

are optimized on the linguistic head, and instead start investigating the use of deeper contextual knowledge and reasoning to perform better on the long tail.

2 The Efficiency of Human Language

Textual documents are surrounded by rich context that is typically leveraged by humans but largely ignored by machines. Ambiguity of language resolves using this context: people optimize their communication to convey maximum information with minimum effort or text given the specific situation. Regardless of the genre (newswire, tweets, fiction, etc.), the Gricean maxim of quantity [4] dictates that an author makes assumptions about the familiarity of the reader with the events and entities that are described in a document at the time of publishing. The author uses this to formulate a message in the most efficient and scarce, yet understandable way. The reader is expected to adequately disambiguate forms and fill in the gaps with presumed knowledge from the current world.

For example, when reading a news item, human readers are aware on which date it was published, the events that occurred around that date, which entities are in the news and recent news articles. Machines, on the contrary, are deprived of such context and expectations. They usually have to deal with individual, isolated news articles, and need to establish identity solely on the basis of a single document in relation to dominant entities in the available resources. To overcome this, we need to build algorithms that can fill contextual knowledge gaps similar to humans with the right assumptions on the familiarity of the entities within a given context. We expect that these considerations are particularly relevant for long-tail entities that are only known within very specific contextual conditions.

3 Types of Knowledge

In [9], four types of contextual knowledge are defined that are essential for humans to interpret text. Here, we relate these four categories to the EL task.
Intratextual knowledge is any knowledge extracted from the text of a document, concerning entity mentions, other word types (e.g. nouns, verbs), and their order and structure in the document. It relates to framing new and given information and notions such as topic and focus. Central entities in the discourse are referred to differently than peripheral ones. Intratextual knowledge is prominent in EL systems: surrounding words (word clouds) [2], entity order [7], coreference [8], substrings [15], abbreviations [7], word senses [10], word relations [1].
Extratextual knowledge concerns any entity-oriented knowledge, found outside the document in (un)structured knowledge bases. Extratextual knowledge can be episodic or conceptual. The former is the knowledge about a concrete entity: its labels, relation to other entities and other facts or experiences. Conceptual knowledge refers to the expectations and knowledge gaps that are filled by an abstract model (i.e. ontology), representing relations between types of entities. Customary extratextual knowledge includes: entity-to-entity links [15, 10, 7, 8], entity labels [15, 10, 7, 8], semantic types [8, 7], and textual descriptions [2].

Circumtextual knowledge - Documents are published at a specific time and location, written by a specific author, and released by a certain publisher. We refer to these prefix and suffix items around the text as *circumtextual knowledge*.
Intertextual knowledge - Documents are not self-contained and rely on *intertextual (cross-document) knowledge* distilled by the reader from related documents. They are published in a stream of information and news, assuming knowledge about preceding related documents, which typically share the same topic and community, and may be published around the same time and location. Early documents that introduce a topic typically make more explicit reference than those published later on when both the event and the topic have evolved.[4]

To the best of our knowledge, circumtextual and intertextual knowledge are systematically neglected in current systems. Then it is no surprise that they fail to handle a case such as the *Hobbs murder* that is presented in the next section.

4 Entity Linking in the Long Tail

In the local news article titled "Hobbs man arrested in connection to nephew's murder",[5] a murder is reported that happened in Hobbs, New Mexico in 2016. It involves two long-tail entities: the killer Michael Johnson and its victim Zachariah Fields. Both entities have no representation in Wikipedia, as they are not well-known outside the context of this murder.

Current EL systems perform poorly on this document. For instance, Babelfy [10] links "Michael Johnson" to a retired American sprinter, "Johnson" to an American president, and "Zachariah" to a long-deceased religious clergyman and author from the 19th century. Not only are these interpretations incorrect, they are also highly incoherent from a human perspective: a retired sprinter, a 19th century religious author, and an ex-president are all identified in an article reporting a local murder in New Mexico in 2016.

What makes these interpretations silly to humans, but optimal to EL systems, is the different notion of coherence. Roughly, entity linkers define coherence via a probabilistic optimization over entity and word associations, resulting in interpretations that neither share context among themselves, nor with the document. Unlike machines, people employ rigorous contextual reasoning over time, location, topic, and other circumtextual knowledge about the article. Time would help to decide against the 19th century author as a victim in 2016. Similarly for location and topic: none of the system interpretations is related to Hobbs, New Mexico, or to any violent event. As systems do not use circumtextual knowledge, they have no human-like mechanisms to decide on improbable interpretations.

In addition, this document is not self-contained; it provides an update regarding an event that happened and was already reported on earlier. In such cases, its interpretation might benefit from (or even depend on) focused machine reading of earlier documents covering this topic. This is very natural for humans; still, current systems lack ways to obtain and integrate intertextual knowledge.

[4] Compare the use of hashtags in Twitter streams once an event becomes trending.
[5] `https://goo.gl/Gms7IQ` Last visited: 18 April 2017

5 Going Forward

While authors assume that their readers possess knowledge from all four categories, we observe that intertextual and circumtextual knowledge are systematically neglected by current entity linkers. This lack of knowledge prevents EL systems to resolve (long-tail) entities that rely on related documents or on contextual awareness regarding publication time, location, topic, author, etc.

Firstly, resolving documents with context-specific, long-tail entities with the current knowledge is a matter of luck and number-crunching and, considering the vast ambiguity of surface forms, extremely challenging in practice. Systems thus need to incorporate circumtextual and intertextual reasoning to compute the coherence of an interpretation and to discourage interpretations that share no context among themselves or the document. How to best employ this knowledge for EL, and combine it with existing probabilistic methods is to be investigated. One could, for instance, model world expectations based on circumtextual aspects, and inspect if the proposed interpretations in text match these expectations.[6]

Secondly, the knowledge from each type should be used systematically and according to its relevance. For instance, the intratextual knowledge is crucial for fictional stories, but not for epitaphs. Even more importantly, approaches that rely on rich and systematic knowledge would be suitable to detect *hunger for knowledge*,[7] i.e. decide that the accessible knowledge is too scarce for making a well-informed decision and assume that entities may be new or unknown to the knowledge base.[8] Once detected, the system has to decide which strategies can be applied to satisfy the hunger, i.e. obtain this missing knowledge. An example for such a strategy when the intratextual information is incomplete is to find information on the same event from related, more explicit documents (cf. [11]).

Finally, the properties of the EL datasets largely determine the types and depths of knowledge needed. Current datasets tend to focus on head entities, so even context-neutral features such as PageRank popularity of an entity lead to F_1 scores of over 60% [14]. Moreover, these datasets often consist of self-contained documents (e.g. Wikipedia samples or sports results), thus consistently removing the need for cross-document knowledge. Instead, systems should be evaluated on long-tail entities by either introducing new long-tail dataset(s) that deliberately contain entities with low dominance and high ambiguity; or by focusing on the few "hard" cases in current datasets. The long-tail cases should also be placed in a broader perspective of the spatio-temporal context by providing documents as a stream of information over time and related to the specific location, thus not only adding documents referring to *Ronaldos* other than the most popular ones, but also the relevant topical stream of documents involving these entities.

We are working on developing datasets that represent long-tail entities better. But considering the complexity and the richness of the long-tail phenomena, this will ultimately need to be a research community effort.

[6] For instance, we would not expect that a 19th century person is still alive in 2016.

[7] We base our concept of hunger for knowledge on [5].

[8] Most probabilistic systems also decide if an entity is new to a knowledge base. However, they set confidence thresholds to circumvent the complexity of this decision.

Acknowledgments

The research for this paper was supported by the Netherlands Organisation for Scientific Research (NWO) via the Spinoza fund and the CLARIAH-CORE project. We thank Stefan Schlobach, Frank van Harmelen, Eduard Hovy, and the reviewers for their ideas and input.

References

1. Cheng, X., Roth, D.: Relational inference for wikification. Urbana 51(61801), 16–58 (2013)
2. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of SEMANTiCS. pp. 121–124. ACM (2013)
3. van Erp, M., Mendes, P., Paulheim, H., Ilievski, F., Plu, J., Rizzo, G., Waitelonis, J.: Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In: LREC. ELRA (2016)
4. Grice, P.: Logic and conversation. In: Cole, P., Morgan, J. (eds.) Syntax and semantics. vol. 3, pp. 41–58. Academic Press, New York (1975)
5. Hovy, E.: Filling the long tail (2016), `https://goo.gl/ieulpF`, Keynote slides from the "Looking at the Long Tail" workshop, 24th June 2016, VU Amsterdam
6. Ilievski, F., Postma, M., Vossen, P.: Semantic overfitting: what world do we consider when evaluating disambiguation of text? In: proceedings of COLING (2016)
7. Ilievski, F., Rizzo, G., van Erp, M., Plu, J., Troncy, R.: Context-enhanced adaptive entity linking. LREC 2016 (2016)
8. Ling, X., Singh, S., Weld, D.S.: Design challenges for entity linking. TACL 3, 315–328 (2015)
9. MacLachlan, G., Reid, I.: Framing and interpretation (1994)
10. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. TACL 2, 231–244 (2014)
11. Narasimhan, K., Yala, A., Barzilay, R.: Improving information extraction by acquiring external evidence with reinforcement learning (2016)
12. Nguyen, T.H., Fauceglia, N., Muro, M.R., Hassanzadeh, O., Gliozzo, A.M., Sadoghi, M.: Joint learning of local and global features for entity linking via neural networks. In: proceedings of COLING (2016)
13. Postma, M., Izquierdo, R., Agirre, E., Rigau, G., Vossen, P.: Addressing the mfs bias in wsd systems. In: Proceedings of LREC 2016. ELRA, Paris, France (2016)
14. Tristram, F., Walter, S., Cimiano, P., Unger, C.: Weasel: a machine learning based approach to entity linking combining different features. In: Proceedings of 3th International Workshop on NLP and DBpedia, ISWC 2015 (2015)
15. Usbeck, R., Ngomo, A.C.N., Röder, M., Gerber, D., Coelho, S.A., Auer, S., Both, A.: Agdistis-graph-based disambiguation of named entities using linked data. In: ISWC. pp. 457–471. Springer (2014)
16. Vossen, P., Agerri, R., Aldabe, I., Cybulska, A., van Erp, M., Fokkens, A., Laparra, E., Minard, A.L., Aprosio, A.P., Rigau, G., Rospocher, M., Segers, R.: Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. Special Issue Knowledge-Based Systems, Elsevier (2016)
17. Zwicklbauer, S., Seifert, C., Granitzer, M.: Doser-a knowledge-base-agnostic framework for entity disambiguation using semantic embeddings. In: ISWC. pp. 182–198. Springer (2016)