Entity Linking: Detecting Entities within Text

Deepak P[1] Sayan Ranu[2]
[1]*IBM Research – India, Bangalore, India*
[2]*Dept. of CS&E, IIT Madras, Chennai, India*
deepak.s.p@in.ibm.com sayan@cse.iitm.ac.in

1. MOTIVATION AND SUMMARY

With unstructured text on the web and social media increasing at a furious pace, it is all the more important to develop techniques that can ease semantic understanding of text data for humans. One of the key tasks in this process is that of entity linking; identifying *mentions* of entities in text. Consider the line that reads *"The Prime Minister came under harsh criticism over the Immigration Act 2014"* Without any additional context, it is not obvious to humans as to who is being talked about. An entity linking technique that has the entity database at its disposal, however, can easily figure out that the mention *Prime Minister* refers to the *Prime Minister of UK* since the mention of *Immigration Act 2014* in the same sentence narrows down the search space from the set of all countries that have Prime Ministers to just UK. Such linking of text documents to entities enables easier understanding for the reader, as well as improved accuracy in automated tasks such as text document clustering, classification and information retrieval.

With the advent of social media, the set of entities that have a presence on the web has increased from just famous places, objects and people, to everyone that has a social media presence, which is to say, virtually the vast majority of human beings. Availability of such a heterogeneous set of entities ranging from those in domain-specific ontologies to social media profiles provides fresh challenges and opportunities for entity linking. In this tutorial, we will cover the set of entity linking techniques that have been proposed in literature over the years, and provide a systematic survey of them with classifications along various dimensions. We will also explore the applicability of entity linking on noisy and short texts, such as those generated in microblogging platforms (ex. Twitter), and elaborate on the new challenges for entity linking that have not quite received enough attention from the scholarly community.

2. TUTORIAL ORGRANIZATION

We propose to organize this as a 1.5 hour tutorial. A brief outline of the tutorial content is as follows:

- **Introduction** (10 minutes)
  - In this segment, we will introduce the task of entity linking with examples as well as technical formalisms. We will motivate the problem and illustrate how entity linking can help in improving traditional learning tasks such as classification and clustering. We will also outline how entity linking differs from closely related tasks such as information extraction and named-entity detection.
- **Considerations in Entity Linking** (25 minutes)
  - We will next introduce the three phases of entity linking, viz., *mention detection*, *candidate discovery* and *entity assignment*. Of these, we will particularly focus on the three criteria that are used in the last phase of entity assignment, i.e., *entity popularity*, *entity-mention similarity* and *document-level coherence*. We will outline the measures that are often used in quantifying each of these notions; for example, entity popularity is often quantified using anchor texts [3], whereas entity-mention similarity is estimated using text similarity metrics [1]. Document-level coherence of entities, on the other hand, is a set-level property and is estimated using graph-mining techniques such as in AIDA [8].
- **Classification of Entity-Linking Techniques** (15 minutes)
  - Entity Linking methods may be classified based on various attributes; in this section, we will analyze entity linking techniques with respect to two major attributes, those pertaining to *usage of supervision* and *document length*. Along the first dimension, we will outline the usage of supervision in techniques such as those in [5] and [7] and the approaches followed by the more popular paradigm of unsupervised entity linking [4, 3]. Most entity linking techniques focus on document-type articles; in this context, we will also delve into techniques that deal with short texts [2] and tweets [6].
- **Evaluation of Entity Linking** (10 minutes)
  - Entity Linking techniques are evaluated using common IR-based metrics such as precision, MAP, MRR and NDCG when ranked lists are output by the techniques [1]. On the other hand, if the entities are returned as sets, set-based evaluation metrics such as recall and F-measure are used. We will introduce these metrics and provide intuitions on which metrics are suitable for various scenarios.

[1]http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html

The 20th International Conference on Management of Data (COMAD), 17th-19th Dec 2014 at Hyderabad, India.

- **Resources for Entity Linking** (10 minutes)
  - Towards motivating the audience to consider entity linking as a field of study and/or exploration, we will outline the various resources that are readily available on the web. These include entity repositories such as Wikipedia[2], Yago[3] as well as numerous text collections. We will also include pointers to entity linking systems that can be accessed on the web.
- **Challenges in Entity Linking** (10 minutes)
  - In this segment, we will systematically explore challenges that have received limited attention from the scholarly community. These include tasks pertaining to entity linking on new entity datasets (e.g., social media profiles) as well as new kinds of document datasets (e.g., scholarly articles, web search queries etc.). Additionally, we will also spend some time discussing methods by which entity linking techniques can enhance general Information Retrieval.
- **Conclusions and Discussion** (10 minutes)

3. TARGETED AUDIENCE & EXPECTATIONS

This tutorial is targeted towards computer scientists interested in the field of data analytics, which includes graduate students and faculty members from academia as well as industry professionals. The tutorial is organized in a self-contained way and does not assume any particular expertise from the audience. By the end of the tutorial, the goal is to expose the audience to the diverse set of problems arising in entity linking, demonstrate how these problems translate to real life applications, and finally, equip attendees with technical insights on how these problems can be solved.

The tutorial is of interest to the COMAD audience since entity linking from text data is a vibrant and active research area due to the omniprescence of social networks in human lives. The tutorial will survey techniques from top publication venues while maintaining a striking balance between the theoretical concepts and their practical importance.

4. BRIEF BIOGRAPHY

Deepak P: Deepak is a researcher in the Information Management Group at IBM Research - India, Bangalore. He obtained his B.Tech degree from Cochin University, India followed by M.Tech and PhD degrees from IIT Madras, India, all in Computer Science. His current research interests include Similarity Search, Spatio-temporal Data Analytics, Graph Mining, Information Retrieval and Machine Learning. He is a senior member of the ACM and IEEE.

Sayan Ranu: Sayan is an Assistant Professor at IIT Madras. Prior to joining IIT Madras, he was a researcher in the Information Management group at IBM Research - India, Bangalore. He obtained his PhD from University of California, Santa Barbara. His current research interests include spatio-temporal data analytics, graph indexing and mining, and bioinformatics.

[2]http://en.wikipedia.org

[3]http://www.mpi-inf.mpg.de/yago-naga/yago

5. REFERENCES

[1] J. Dalton and L. Dietz. A neighborhood relevance model for entity linking. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pages 149–156. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2013.

[2] P. Ferragina and U. Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM, 2010.

[3] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics, 2011.

[4] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–466. ACM, 2009.

[5] Y. Li, C. Wang, F. Han, J. Han, D. Roth, and X. Yan. Mining evidences for named entity disambiguation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1070–1078. ACM, 2013.

[6] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 563–572. ACM, 2012.

[7] A. Pilz and G. Paaß. Collective search for concept disambiguation. 2012.

[8] M. A. Yosef, J. Hoffart, I. Bordino, M. Spaniol, and G. Weikum. Aida: An online tool for accurate disambiguation of named entities in text and tables. *Proceedings of the VLDB Endowment*, 4(12):1450–1453, 2011.