Named Entity Extraction for Knowledgebase Enhancement

Priya Radhakrishnan
IIIT Hyderabad, India
priya.r@research.iiit.ac.in

Abstract

Past decade witnessed an explosive growth in the amount of unstructured data, especially in the public domain, mainly due to Web 2.0 and social media. This led to the creation of applications, called information extractors, that extract structured information from unstructured data. The extracted information is stored in a Knowledge Base (KB). KB stores facts about entities like name, type and other attributes.

My PhD thesis entitled 'Named Entity Extraction for Knowledgebase Enhancement' deals with information extraction on named entities with the purpose of enhancing a KB. The enhanced KB is in turn used by the information extraction task to refine the extraction process. Thus, KB provides structure and guidance to the extraction task, and gets enhanced by the results of the extraction task. Here we see that the tasks of entity extraction and KB enhancement are mutually dependent and mutually beneficial. Hence in my research I propose methods to enhance both the tasks, in an effort to build a strong and sound named entity extraction system.

Named Entity Extraction, also known as Entity Linking (EL) in scientific literature, is the task of determining the identity of entities mentioned in text. EL helps automatic extraction of structured information about entities from unstructured data, which is stored in the KB. EL consists of Mention Detection and Entity Disambiguation. In my research, I propose methods to enhance mention detection, entity disambiguation and KB enhancement.

Textual content to be processed by information extractor, typically is about multiple named entities. The information extractor has to identify the named entities that are important to the content, a.k.a. salient named entities. My thesis proposes a method to identify salient named entity of text [6]. Salience of a named entity can also be judged by learning how the named entity is semantically related to other named entities mentioned in the document. We propose a method to identify the semantic relations within named entities [4]. The proposed methods help to improve mention detection.

Performance of Named Entity Extraction methods depend on the size and structure of the context of the named entity mention in the text. While bigger size and better structure of context result in improved performance of the Named Entity Extraction, smaller size and poor structure of context result in reduced performance. We propose three different Named Entity Extraction approaches tailored to the varying size and structure of the context in this thesis [7, 1, 3]. The proposed methods perform on par with state-of-the-art methods with improved latency.

Named Entity Extraction approaches that work on lesser context and lower structure, increasingly depend on non-textual signal like global coherence of entities in the KB. Degree of connectivity of an entity in the KB directly affects an EL systems ability to correctly link mentions in text to the entity in KB. This causes many EL systems to perform well for entities well connected to other entities in KB, bringing focus to connection density of KB in EL. We propose ELDEN, an EL system that densifies the KB with co-occurrence statistics from a large text corpus and uses the densified KB to train entity embeddings. Entity similarity measured using these trained entity embeddings results in improved EL [5]. ELDEN outperforms state-of-the-art EL system on benchmark datasets.

Proof of the pudding is in the eating. We demonstrate use of entities stored in KB in an application, an information retrieval task that is improved by use of KB entities. We propose a novel method for enhancing classification performance of research papers into ACM computer science categories using KB entities, namely Wikipedia and Freebase entities [2].

All through my research I have proposed methods of improving Named Entity Extraction using KBs. The approaches model word and entity representations for automatic text understanding using KB. My current and future explorations will also be towards building better representations of entity and document towards achieving this goal.

References

[1] R. Bansal, S. Panem, P. Radhakrishnan, M. Gupta, and V. Varma. Linking entities in #microposts. In M. Rowe, M. Stankovic, and A. Dadzie, editors, *Proceedings of the the 4th Workshop on Making Sense of Microposts co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 7th, 2014.*, pages 71–72. CEUR-WS.org, 2014.

[2] S. Gupta, P. Radhakrishnan, M. Gupta, V. Varma, and U. Gupta. Enhancing categorization of computer science research papers using knowledge bases. In *Proceedings of the First Workshop on Knowledge Graphs and Semantics for Text Retrieval and Analysis (KG4IR 2017) co-located with the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017), Shinjuku, Tokyo, Japan, August 11, 2017.*, pages 38–42, 2017.

[3] P. Radhakrishnan, R. Bansal, M. Gupta, and V. Varma. Exploiting Wikipedia Inlinks for Linking Entities in Queries. In *Proceedings of the First International Workshop on Entity Recognition & Disambiguation*, ERD '14, pages 101–104, New York, NY, USA, 2014. ACM.

[4] P. Radhakrishnan, M. Gupta, and V. Varma. Modeling the evolution of product entities. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 923–926, New York, NY, USA, 2014. ACM.

[5] P. Radhakrishnan, M. Gupta, and V. Varma. Elden: Improved entity linking using densified knowledge graphs. In *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, 2018.

[6] P. Radhakrishnan, G. Jawahar, M. Gupta, and V. Varma. Sneit: Salient named entity identification in tweets. *Computación y Sistemas*, 21(4):665679, 2017.

[7] V. Varma, B. Ghosh, M. Soundararajan, D. Aggarwal, and P. Radhakrishnan. IIIT hyderabad at TAC 2012. In *Proceedings of the Fifth Text Analysis Conference, TAC 2012, Gaithersburg, Maryland, USA, November 5-6, 2012*. NIST, 2012.