Papers for comprehensive viva-voce

Priya Radhakrishnan

Advisor : Dr. Vasudeva Varma
Search and Information Extraction Lab,
International Institute of Information Technology,
Gachibowli, Hyderabad, India 500032

1 Papers selected and their justification:

Towards research in Entity Linking and Knowledgebase Construction, we mainly focus on papers that have appeared in reputed conferences that reflect works in Entity Linking and Knowledgebase Construction. This includes issues like wikification, Named Entity Disambiguation, Entity Extraction and Relation Extraction. We broadly classify 14 papers that we have selected as follows:

1. Entity Linking : Papers 1,4,5 and 6
2. Entity Linking in short texts : Papers 7 and 8
3. Entity Linking Evaluation : Papers 9 and 10
4. Entity Attribute Extraction : Paper 14
5. Knowledgebase Construction : Papers 12 and 13
6. Semantic Search : Papers 2,3 and 11

Table 1: List of selected Papers

SI	Field	Publication	Citation	Reason for selecting
1	Entity Linking	[1]	175	This paper introduces the use of Wikipedia as a resource for automatic keyword extraction and word sense disambiguation. It shows how Wikipedia can be used to achieve state-of-the-art results on both these tasks. The paper also shows how the two methods can be combined into a system able to automatically enrich a text with links to Wikipedia. Given an input document, the system identifies the important concepts in the text and automatically links these concepts to the corresponding Wikipedia pages.

Continued on next page

Table 1 – *Continued from previous page*

SI	**Field**	**Publication**	**Citation**	**Reason for selecting**
2	Semantic Search	[2]	328	The paper proposes Explicit Semantic Analysis (ESA), a novel method that represents the meaning of texts in a high-dimensional space of concepts derived from Wikipedia. It uses machine learning techniques to explicitly represent the meaning of any text as a weighted vector of Wikipedia-based concepts. Assessing the relatedness of texts in this space amounts to comparing the corresponding vectors using conventional metrics (e.g., cosine). Compared with the previous state of the art, using ESA results in substantial improvements in correlation of computed relatedness scores with human judgments.
3	Semantic Search	[3]	380	This paper describes a new technique for obtaining measures of semantic relatedness. It uses Wikipedia to provide structured world knowledge about the terms of interest. The approach uses the hyperlink structure of Wikipedia rather than its category hierarchy or textual content. Evaluation with manually defined measures of semantic relatedness reveals this to be an effective compromise between the ease of computation of the former approach and the accuracy of the latter
4	Entity Linking	[4]	597	This paper describes how to automatically cross-reference documents with Wikipedia. It explains how machine learning can be used to identify significant terms within unstructured text, and enrich it with links to the appropriate Wikipedia articles. This link detector and disambiguator performs very well, with recall and precision of almost 75%. This performance is constant whether the system is evaluated on Wikipedia articles or ”real world” documents.

Continued on next page

Table 1 – *Continued from previous page*

SI	**Field**	**Publication**	**Citation**	**Reason for selecting**
5	Entity Linking	[5]	396	This paper presents a large-scale system for the recognition and semantic disambiguation of named entities based on information extracted from a large encyclopedic collection and Web search results. It describes in detail the disambiguation paradigm employed and the information extraction process from Wikipedia. Through a process of maximizing the agreement between the contextual information extracted from Wikipedia and the context of a document, as well as the agreement among the category tags associated with the candidate entities, the implemented system shows high disambiguation accuracy on both news stories and Wikipedia articles.
6	Entity Linking	[6]	204	This paper presents a system to link entity mentions on Web pages to entities in Wikipedia. This paper proposes a general collective disambiguation approach. On the premise that coherent documents refer to entities from one or a few related topics or domains, the authors propose formulations for the trade-off between local spot-to-entity compatibility and measures of global coherence between entities. The proposed solution is based on local hill-climbing, rounding integer linear programs, and pre-clustering entities followed by local optimization within clusters. In experiments involving over a hundred manually-annotated Web pages and tens of thousands of entity mentions, the approach significantly outperforms other existing algorithms.
7	Entity Linking in short texts	[7]	40	This work uses Wikipedia's anchor text to page mapping to address the problem of cross-referencing text fragments with Wikipedia pages. This way synonymy and polysemy issues are resolved accurately and efficiently.

Continued on next page

Table 1 – *Continued from previous page*

SI	**Field**	**Publication**	**Citation**	**Reason for selecting**
8	Entity Linking in short texts	[8]	29	This paper propose a solution to the problem of determining what a microblog post (tweet) is about through semantic linking. It adds semantics to tweets by automatically identifying concepts that are semantically related to it and generating links to the corresponding Wikipedia articles. The identified concepts can subsequently be used for, e.g., social media mining, thereby reducing the need for manual inspection and selection.
9	Entity Linking Evaluation	[9]	29	This paper presents a benchmarking framework for fair and exhaustive comparison of entity-annotation systems. The framework is based upon the definition of a set of problems related to the entity-annotation task, a set of measures to evaluate systems performance, and a systematic comparative evaluation involving all publicly available data-sets, containing texts of various types such as news, tweets and Web pages.
10	Entity Linking Evaluation	[10]	12	This paper re-implements three seminal Named Entity Linking (NEL) systems and presents a detailed evaluation of search strategies. The results are systematically compared on standard data sets. The results establish that co-reference and acronym handling lead to substantial improvement, and search strategies account for much of the variation between systems.
11	Semantic Search	[11]	175	This work presents experiments on using Wikipedia for computing semantic relatedness and compares it to WordNet on various benchmarking datasets. Existing relatedness measures perform better using Wikipedia than a baseline given by Google counts. It also shows that Wikipedia outperforms WordNet when applied to the largest available dataset designed for that purpose. The best results on this dataset are obtained by integrating Google, WordNet and Wikipedia based measures.

Continued on next page

Table 1 – *Continued from previous page*

SI	**Field**	**Publication**	**Citation**	**Reason for selecting**
12	Knowlege Base Construction	[12]	121	In this paper authors compare the two graphs in Wikipedia (i) the article graph, and (ii) the category graph. In a graph theoretic analysis of the category graph, the authors show that Wikipedia Category Graph is a scale-free, small world graph like other well-known lexical semantic networks.
13	Knowlege Base Construction	[13]	1296	This paper describes YAGO. YAGO is a light-weight and extensible ontology with high coverage and quality. YAGO contains more than 1 million entities and 5 million facts. This includes the Is-A hierarchy as well as non-taxonomic relations between entities (such as HASONEPRIZE). The facts were automatically extracted from Wikipedia and unified with WordNet, using a carefully designed combination of rule-based and heuristic method described in this paper.
14	Entity Attribute Extraction	[14]	24	This paper describes extracting attribute and value pairs from textual product descriptions. The goal is to augment databases of products by representing each product as a set of attribute-value pairs. Such a representation is beneficial for tasks where treating the product as a set of attribute-value pairs is more useful than as an atomic entity. Examples of such applications include demand forecasting, assortment optimization, product recommendations, and assortment comparison across retailers and manufacturers. The problem is formulated as a classification problem and solved using semi-supervised learning algorithms.

2 Justification for selecting these papers:

Some relevant papers to develop background knowledge for starting research on Entity Linking specially focusing on Knowledgebase construction and Semantic search.

The papers are selected to get the complete overview and motivation to do research in **Entity Linking (EL)**. Entity Linking is the task of linking textual entity mentions to entries in a knowledge base that contain relevant information regarding the entities. Wikipedia (or Wikipedia derived KB like

YAGO or Freebase) is typically used as the KB in EL studies. The performance of an EL system depends on how well the KB understands entities, which requires higher quantity and quality of entity representation in KB. Thus half of the selected papers deal with Entity Linking - in documents and text, the later half deals with Knowledgebase construction and maintanance. As the research focus is mainly on challenges due to name variations and entity ambiguity. And thus papers specific to these challenges are selected. After the study of these papers current research issues related to Knowledgebase enhancements are identified.

References

1. Mihalcea, R., Csomai, A.: Wikify!: Linking documents to encyclopedic knowledge. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management. CIKM '07, New York, NY, USA, ACM (2007) 233–242
2. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th international joint conference on Artifical intelligence. IJCAI'07, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2007) 1606–1611
3. Milne, D., Witten, I.H.: An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In: In Proceedings of AAAI 2008. (2008)
4. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. CIKM '08, New York, NY, USA, ACM (2008) 509–518
5. Cucerzan, S.: Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In: Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Association for Computational Linguistics (Jun 2007) 708–716
6. Kulkarni, S., Singh, A., Ramakrishnan, G., Chakrabarti, S.: Collective annotation of wikipedia entities in web text. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '09, New York, NY, USA, ACM (2009) 457–466
7. Ferragina, P., Scaiella, U.: TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities). In: Proc. of the 19^{th} ACM Intl. Conf. on Information and Knowledge Management (CIKM). (2010) 1625–1628
8. Meij, E., Weerkamp, W., de Rijke, M.: Adding Semantics to Microblog Posts. In: Proc. of the 5^{th} ACM Intl. Conf. on Web Search and Data Mining (WSDM), ACM (2012) 563–572
9. Cornolti, M., Ferragina, P., Ciaramita, M.: A framework for benchmarking entity-annotation systems. In: Proceedings of the 22Nd International Conference on World Wide Web. WWW '13, Republic and Canton of Geneva, Switzerland, International World Wide Web Conferences Steering Committee (2013) 249–260
10. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J.R.: Evaluating entity linking with wikipedia. Artif. Intell. **194** (January 2013) 130–150

11. Strube, M., Ponzetto, S.P.: Wikirelate! computing semantic relatedness using wikipedia. In: proceedings of the 21st national conference on Artificial intelligence - Volume 2. AAAI'06, AAAI Press (2006) 1419–1424
12. Zesch, T., Gurevych, I.: Analysis of the wikipedia category graph for nlp applications. In: Proceedings of the TextGraphs-2 Workshop (NAACL-HLT), Rochester, Association for Computational Linguistics (April 2007) 1–8
13. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web. WWW '07, New York, NY, USA, ACM (2007) 697–706
14. Ghani, R., Probst, K., Liu, Y., Krema, M., Fano, A.: Text mining for product attribute extraction. SIGKDD Explorations **1** (2006) 41–48