Wikipedia as a Resource for Text Analysis and Retrieval

Marius Paşca
Google
1600 Amphitheatre Parkway
Mountain View, California 94043
mars@google.com

1 Tutorial Description

As a counterpart to expert-created knowledge resources such as WordNet or Cyc, non-expert users may collaboratively create large resources of unstructured or semi-structured knowledge, a leading representative of which is Wikipedia. Collectively, articles within Wikipedia form an easily-editable collection, reflecting an ever-growing number of topics of interest to Web users.

This tutorial examines the characteristics of Wikipedia relative to other human-curated resources of knowledge; and the role of Wikipedia and resources derived from it in text analysis and in enhancing information retrieval. Applicable text analysis tasks include coreference resolution (Ratinov and Roth, 2012), word sense and entity disambiguation (Ganea and Hofmann, 2017). More prominently, they include information extraction (Zhu et al., 2019). In information retrieval, a better understanding of the structure and meaning of queries (Hu et al., 2009; Pantel and Fuxman, 2011; Tan et al., 2017) helps in matching queries against documents (Ensan and Bagheri, 2017), clustering search results (Scaiella et al., 2012), answer (Chen et al., 2017) and entity retrieval (Ma et al., 2018) and retrieving knowledge panels for queries asking about popular entities.

2 Outline

1. Human-curated resources
   (a) Expert resources
   (b) Collaborative, non-expert resources
   (c) Hybrid resources
2. Knowledge within Wikipedia
   (a) Articles, infoboxes, links, categories
   (b) Resources derived from Wikipedia
3. Role in text analysis
   (a) Information extraction
   (b) Beyond information extraction
4. Role in information retrieval
   (a) Query and document analysis
   (b) Retrieval and ranking

A copy will be at http://tinyurl.com/acl19wi.

3 Presenter

Marius Paşca is a research scientist at Google in Mountain View, California. He graduated with a Ph.D. in Computer Science from Southern Methodist University in Dallas, Texas and an M.Sc. in Computer Science from Joseph Fourier University in Grenoble, France. Current research interests include factual information extraction from unstructured text and natural-language matching functions for information retrieval.

References

D. Chen, A. Fisch, J. Weston, and A. Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *ACL 2017*.

F. Ensan and E. Bagheri. 2017. Document retrieval model through semantic linking. In *WSDM 2017*.

O. Ganea and T. Hofmann. 2017. Deep joint entity disambiguation with local neural attention. In *EMNLP 2017*.

J. Hu, G. Wang, F. Lochovsky, J. Sun, and Z. Chen. 2009. Understanding user's query intent with Wikipedia. In *WWW 2009*.

D. Ma, Y. Chen, K. Chang, and X. Du. 2018. Leveraging fine-grained Wikipedia categories for entity search. In *WWW 2018*.

P. Pantel and A. Fuxman. 2011. Jigs and lures: Associating web queries with structured entities. In *ACL 2011*.

L. Ratinov and D. Roth. 2012. Learning-based multi-sieve co-reference resolution with knowledge. In *EMNLP-CoNLL 2012*.

U. Scaiella, P. Ferragina, A. Marino, and M. Ciaramita. 2012. Topical clustering of search results. In *WSDM 2012*.

C. Tan, F. Wei, P. Ren, W. Lv, and M. Zhou. 2017. Entity linking for queries by searching Wikipedia sentences. In *EMNLP 2017*.

Q. Zhu, X. Ren, J. Shang, Y. Zhang, A. El-Kishky, and J. Han. 2019. Integrating local context and global cohesiveness for open information extraction. In *WSDM 2019*.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, page 24
Florence, Italy, July 28 - August 2, 2019.