DOCTORAL ABSTRACT

Approaches for Enriching and Improving Textual Knowledge Bases

Besnik Fetahu
L3S Research Center, Leibniz University of Hannover
fetahu@L3S.uni-hannover.de

Abstract

Verifiability is one of the core editing principles in Wikipedia, where editors are encouraged to provide *citations* for the added statements. Statements can be any arbitrary piece of text, ranging from a sentence up to a paragraph. However, in many cases, citations are either outdated, missing, or link to non-existing references (e.g. dead URL, moved content etc.). In total, 20% of the cases such citations refer to *news* articles and represent the second most cited source. Even in cases where citations are provided, there are no explicit indicators for the span of a citation for a given piece of text. In addition to issues related with the verifiability principle, many Wikipedia entity pages are incomplete, with relevant information that is already available in online news sources missing. Even for the already existing citations, there is often a delay between the news publication time and the reference time.

In this thesis, we address the aforementioned issues and propose automated approaches that enforce the *verifiability* principle in Wikipedia, and suggest relevant and missing news references for further enriching Wikipedia entity pages. To this end we make the following contributions as part of this thesis [1, 2, 3, 4]:

- *Citation recommendation* – we address the problem of finding and updating news citations for statements in Wikipedia entity pages [3]. We propose a two-stage approach for this problem. First, we classify each statement whether it requires a news citation or citations from other categories (e.g. web, book, journal, etc.). Second, for statements that require a news citation, we formalize three properties of what makes a good citation, namely: (i) the citation should entail the Wikipedia statement, (ii) the statement should be central to the citation, and (iii) the citation should be from an authoritative source. We combine standard information retrieval techniques, where we use the statement to query a news collection, and build classification models based on the three properties to determine the most appropriate citation.
- *Citation span* – from the already existing citations in Wikipedia entity pages and the ones we recommend in our first problem, we propose an automated approach which determines the span of such citations [4]. We approach this problem by classifying which textual fragments in a paragraph are covered or hold true given a citation. We propose a sequence classification approach where for a paragraph and a citation, we determine the citation span at a fine-grained level.

- *News suggestion* – to account for the ever evolving nature of Wikipedia entities, with relevant information published on a daily basis in news articles, we propose a two-stage supervised approach for this problem [1]. First, we suggest news articles to Wikipedia entities (article-entity placement) relying on a rich set of features which take into account the *salience* and *relative authority* of entities, and the *novelty* of news articles to entity pages. Second, we determine the exact section in the entity page for the input article (article-section placement) guided by class-based section templates.

We perform extensive evaluation with real-world datasets, on news collections with more than 20 million news articles, and on the entire set of english Wikipedia entity pages. Our approaches perform with high accuracy on the three problems we address and show superior performance when compared to existing baselines and state of the art approaches.

The thesis and accompanying material is available at `http://l3s.de/~fetahu`.

References

[1] Besnik Fetahu, Katja Markert, and Avishek Anand. Automated news suggestions for populating wikipedia entity pages. In James Bailey, Alistair Moffat, Charu C. Aggarwal, Maarten de Rijke, Ravi Kumar, Vanessa Murdock, Timos K. Sellis, and Jeffrey Xu Yu, editors, *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 323–332. ACM, 2015.

[2] Besnik Fetahu, Abhijit Anand, and Avishek Anand. How much is wikipedia lagging behind news? In David De Roure, Pete Burnap, and Susan Halford, editors, *Proceedings of the ACM Web Science Conference, WebSci 2015, Oxford, United Kingdom, June 28 - July 1, 2015*, pages 28:1–28:9. ACM, 2015.

[3] Besnik Fetahu, Katja Markert, Wolfgang Nejdl, and Avishek Anand. Finding news citations for wikipedia. In Snehasis Mukhopadhyay, ChengXiang Zhai, Elisa Bertino, Fabio Crestani, Javed Mostafa, Jie Tang, Luo Si, Xiaofang Zhou, Yi Chang, Yunyao Li, and Parikshit Sondhi, editors, *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 337–346. ACM, 2016.

[4] Besnik Fetahu, Katja Markert, and Avishek Anand. Fine grained citation span for references in wikipedia. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, Copenhagen, Denmark, September 7-11, 2017*, 2017.