Knowledge Extraction for Hybrid Question Answering

Ricardo Usbeck

May 18, 2017

Since the proposal of hypertext by Tim-Berners Lee to his employer CERN on March 12, 1989[1] the World Wide Web has grown to more than one billion Web pages and still grows.[2] With the later proposed Semantic Web vision [1], Lee et al. suggested an extension of the existing (Document) Web to allow better reuse, sharing and understanding of data.

Both the Document Web and the Web of Data (which is the current implementation of the Semantic Web) grow continuously. This is a mixed blessing, as the two forms of the Web grow concurrently and most commonly contain different pieces of information. Modern information systems must thus bridge a *Semantic Gap* to allow a holistic and unified access to information about a particular information independent of the representation of the data. One way to bridge the gap between the two forms of the Web is the extraction of structured data, i.e., RDF, from the growing amount of unstructured[3] and semi-structured information (e.g., tables and XML) on the Document Web.

While extracting structured data from unstructured data allows the development of powerful information system, it requires high-quality and scalable knowledge extraction frameworks to lead to useful results. The dire need for such approaches has led to the development of a multitude of annotation frameworks and tools. However, most of these approaches are not evaluated on the same datasets or using the same measures. The resulting *Evaluation Gap* needs to be tackled by a concise evaluation framework to foster fine-grained and uniform evaluations of annotation tools and frameworks over any KBs.

Moreover, with the constant growth of data and the ongoing decentralization of knowledge, intuitive ways for non-experts to access the generated data are required. Humans adapted their search behavior to current Web data by access paradigms such as keyword search so as to retrieve high-quality results. Hence, most Web users only expect Web documents in return [2]. However, humans think and most commonly express their information needs in their natural language rather than using keyword phrases [12]. Answering complex information needs often requires the combination of knowledge from

[1] `http://www.w3.org/People/Berners-Lee/Longer.html`

[2] `http://www.internetlivestats.com/total-number-of-websites/`

[3] Note, that unstructured data stands for any type of textual information like news, blogs or tweets.

various, differently structured data sources. Thus, we observe an *Information Gap* between natural-language questions and current keyword-based search paradigms, which in addition do not make use of the available structured and unstructured data sources. Question Answering (QA) systems provide an easy and efficient way to bridge this gap by allowing to query data via natural language, thus reducing (1) a possible loss of precision and (2) potential loss of time while reformulating the search intention to transform it into a machine-readable way. Furthermore, QA systems enable answering natural language queries with concise results instead of links to verbose Web documents. Additionally, they allow as well as encourage the access to and the combination of knowledge from heterogeneous knowledge bases (KBs) within one answer.

Consequently, three main research gaps are considered and addressed in this work:

1. First, addressing the Semantic Gap between the unstructured Document Web and the Semantic Gap requires the development of scalable and accurate approaches for the extraction of structured data in RDF [6]. This research challenge is addressed by several approaches within this thesis. This thesis presents CETUS [8], an approach for recognizing entity types to populate RDF KBs. Furthermore, our knowledge base-agnostic disambiguation framework AGDISTIS [9] can efficiently detect the correct URIs for a given set of named entities. Additionally, we introduce REX [3], a Web-scale framework for RDF extraction from semi-structured (i.e., templated) websites which makes use of the semantics of the reference knowledge based to check the extracted data.

2. The ongoing research on closing the Semantic Gap has already yielded a large number of annotation tools and frameworks. However, these approaches are currently still hard to compare since the published evaluation results are calculated on diverse datasets and evaluated based on different measures. On the other hand, the issue of comparability of results is not to be regarded as being intrinsic to the annotation task. Indeed, it is now well established that scientists spend between 60% and 80% of their time preparing data for experiments [4, 5, 7]. Data preparation being such a tedious problem in the annotation domain is mostly due to the different formats of the gold standards as well as the different data representations across reference datasets. We tackle the resulting *Evaluation Gap* in two ways: First, we introduce a collection of three novel datasets, dubbed N^3, to leverage the possibility of optimizing NER and NED algorithms via Linked Data and to ensure a maximal interoperability to overcome the need for corpus-specific parsers. Second, we present GERBIL [11], an evaluation framework for semantic entity annotation. The rationale behind our framework is to provide developers, end users and researchers with easy-to-use interfaces that allow for the agile, fine-grained and uniform evaluation of annotation tools and frameworks on multiple datasets.

3. The decentral architecture behind the Web has led to pieces of information being distributed across data sources with varying structure. Moreover, the increasing the demand for natural-language interfaces[4] as depicted by current mobile applica-

[4] `http://www.forbes.com/sites/jaysondemers/2015/11/10/the-fundamental-guide-to-seo-in-2016`

tions requires systems to deeply understand the underlying user information need. In conclusion, the natural language interface for asking questions requires a *hybrid* approach to data usage, i.e., simultaneously performing a search on full-texts and semantic KBs To close the Information Gap, this thesis presents HAWK [10], a novel entity search approach developed for hybrid QA based on combining structured RDF and unstructured full-text data sources.

Bibliographic Data

Title	Knowledge Extraction for Hybrid Question Answering
Author	Ricardo Usbeck
Supervisors	Prof. Dr.-Ing. habil. Klaus-Peter Fähnrich and Dr. Axel-Cyrille Ngonga Ngomo
Institution	Leipzig University, Faculty for Mathematics and Computer Science
Time frame	January 2013 - March 2016

References

[1] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Sci. Amer.*, 284(5):34–43, May 2001.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *7th International World Wide Web Conference*, 1998.

[3] L. Bühmann, R. Usbeck, A.-C. Ngonga Ngomo, M. Saleem, A. Both, V. Crescenzi, P. Merialdo, and D. Qiu. Web-Scale Extension of RDF Knowledge Bases from Templated Websites. In *International Semantic Web Conference (ISWC)*, pages 66–81, 2014.

[4] Y. Gil. Semantic challenges in getting work done, 2014. Invited Talk at ISWC.

[5] P. Jermyn, M. Dixon, and B. J. Read. Preparing clean views of data for data mining. *ERCIM Work. on Database Res*, 1999.

[6] F. Manola and E. Miller. RDF Primer. Technical report, W3C, http://www.w3.org/TR/rdf-primer/, Feb. 2004.

[7] R. D. Peng. Reproducible research in computational science. *Science (New York, Ny)*, 2011.

[8] M. Röder, R. Usbeck, R. Speck, and A.-C. Ngonga Ngomo. CETUS – A Baseline Approach to Type Extraction. In *1st Open Knowledge Extraction Challenge at International Semantic Web Conference*, 2015.

[9] R. Usbeck, A.-C. Ngonga Ngomo, L. Bühmann, M. Röder, D. Gerber, S. Athaide Coelho, S. Auer, and A. Both. AGDISTIS - Graph-Based Disambiguation of Named Entities Using Linked Data. In *International Semantic Web Conference (ISWC)*, pages 457–471, 2014.

[10] R. Usbeck, A.-C. Ngonga Ngomo, L. Bühmann, and C.-t. Unger. HAWK – Hybrid Question Answering Using Linked Data. In *Extended Semantic Web Conference*, pages 353–368, 2015.

[11] R. Usbeck, M. Röder, A.-C. Ngonga Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann. GERBIL – General Entity Annotation Benchmark Framework. In *24th International Conference on World Wide Web*, 2015.

[12] W. A. Woods. Progress in natural language understanding: an application to lunar geology. In *International computer conference and exposition*, pages 441–450. ACM, 1973.