Round-trip semantics with Sztakipedia and DBpedia Spotlight

Mihály Héder
MTA SZTAKI
Victor Hugo 18-22
Budapest, Hungary
mihaly.heder@sztaki.hu

Pablo N. Mendes
Web Based Systems Group
Freie Universität Berlin
pablo.mendes@fu-berlin.de

ABSTRACT

We describe a tool kit to support a knowledge-enhancement cycle on the Web. In the first step, structured data which is extracted from Wikipedia is used to construct automatic content enhancement engines. Those engines can be used to interconnect knowledge in structured and unstructured information sources on the Web, including Wikipedia itself. Sztakipedia-toolbar is a MediaWiki user script which brings DBpedia Spotlight and other kinds of machine intelligence into the Wiki editor interface to provide enhancement suggestions to the user. The suggestions offered by the tool focus on complementing knowledge and increasing the availability of structured data on Wikipedia. This will, in turn, increase the available information for the content enhancement engines themselves, completing a virtuous cycle of knowledge enhancement.

A 90 seconds long screencast instroduces the system on youtube: `http://www.youtube.com/watch?v=8VW0TrvXpl4`. For those who are interested in more details there is an other 4 minutes long video: `http://www.youtube.com/watch?v=cLqe-D0qKCM`.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems; H.5.2 [**Information Systems Applications**]: User Interfaces; H.4 [**Information Systems Applications**]: Hypertext/Hypermedia

General Terms

Wikipedia, Semantic Web, Natural Language Processing

Keywords

Sztakipedia, DBpedia Spotlight

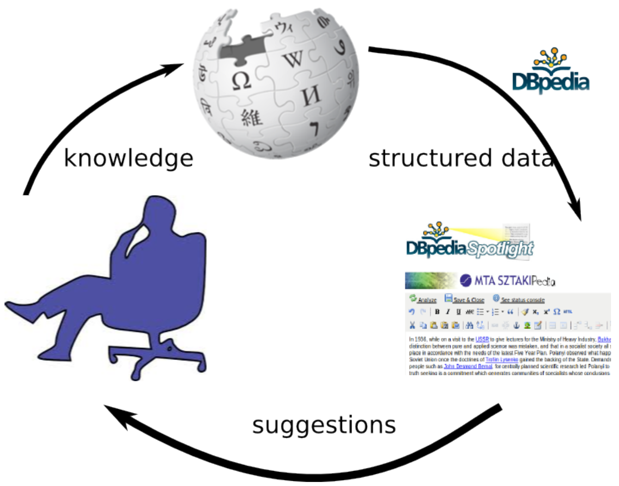
1. INTRODUCTION

Wikipedia is an important collection of the encyclopaedic knowledge of mankind. It contains millions of interlinked articles organized in categories and commonly containing structured summaries such as tables and infoboxes. With such features, Wikipedia presents an unprecedented opportunity for projects which aim at gathering machine representation of human knowledge. One of the most prominent among those projects is DBpedia [1]. The DBpedia project extracts structured information from Wikipedia editions in 97 different languages and combines this information into a large free multilingual knowledge base. Furthermore, for 15 languages it maps facts from the infobox templates into a single consistent ontology. The DBpedia Ontology organizes the knowledge on Wikipedia according to a hierarchy of 320 classes and 1,650 different properties. Mappings between Wikipedia infoboxes and the DBpedia Ontology are community-generated, and allow a more homogenized view of the knowledge e.g. mapping multiple spellings of properties to one canonical name, or recognizing that many infobox types describe the same entity type. DBpedia provides a different view of the information in Wikipedia, allowing database-style queries for batch information processing or for knowledge reuse in other applications.

DBpedia Spotlight [4] is a tool that reuses the knowledge in DBpedia and Wikipedia to enable the automatic semantic annotation of text. The tool inspects natural language text and, based on its content, identifies potential links to DBpedia – and consequently Wikipedia – that help to describe or complement the textual content. The structured information associated with the natural language text can then be used, for instance, to help content exploration through faceted browsing, or to enable enhanced information retrieval based on the complementary information from the knowledge base.

Sztakipedia builds upon machine readable knowledge and intends to be an Intelligent Assistant for knowledge editors, primarily for Wikipedia. The system design is based on a requirement survey detailed in [3]. Sztakipedia uses DBpedia data, among other sources, for making suggestions of different kinds, offering pertinent content for editors to add to the documents, e.g. Wikipedia infoboxes, categories and page links. Among the information sources used, we highlight Web search, library catalogs, as well as TF-IDF database and co-occurrence data extracted from Wikipedia. Through the use of Sztakipedia, Wikipedia users can reuse DBpedia data unknowingly when editing articles, through a toolbar from the standard wiki editor interface (Figure 1). The assisted editing of articles can increase the level of interconnection of existing knowledge and potentially enhance the quality of articles on Wikipedia.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.



In 1936, while on a visit to the USSR to give lectures for the Ministry of Heavy Industry, Bukha
distinction between pure and applied science was mistaken, and that in a socialist society all s
place in accordance with the needs of the latest Five Year Plan. Polanyi observed what happ
Soviet Union once the doctrines of Trofim Lysenko gained the backing of the State. Demands
people such as John Desmond Bernal, for centrally planned scientific research led Polanyi to
truth seeking is a commitment which generates communities of specialists whose conclusions

Figure 1: Round-trip Semantics

2. ROUND-TRIP SEMANTICS

Our vision is that Intelligent Systems (IS) based on Wikipedia knowledge can foster a virtuous cycle of semantic enhancement. The editor interface is supported by the knowledge base to create better documents, which in turn increases the quality and size of the knowledge base as a whole. With a better and larger knowledge base, intelligent systems have more data to learn from, and can therefore provide better assistance to the user, creating a positive feedback loop (Figure 1).

The Intelligent System's role is to make suggestions and recommendations to improve the knowledge source's amount and quality of information. This could mean intra-wiki link suggestions, category and label recommendations, semantic relation suggestion, a bibliography for further reading, etc. The IS may be proactive – triggering a suggestion whenever an improvement opportunity is detected – or act upon user request. An important element of this vision is that the user should make most of the decisions about the suggestions. As a consequence the IS must be present online in the editor interface of the user, e.g. as a plugin.

3. FEATURES OF THE SYSTEM

We present Sztakipedia-toolbar, a MediaWiki user script and supporting server modules, that can be easily enabled by any Wikipedia user – currently fully functional only for the English Wikipedia. The toolbar provides access to the following functions:

- Category recommendation. This function is responsible for helping the user in the hard task of choosing the right Wikipedia categories for the article. This function is enabled by a scoped Yahoo BOSS[1] search performed in the background. We search for relevant categories by querying for strings like: "category `some important words of the text`" while restricting results to the domain *en.wikipedia.org/wiki/Category:* or in the appropriate sites in non-english Wikis.

- Infobox recommendation. The implementation of this function is based on document similarity, calculated by the Lucene[2] framework. The articles in a Wikipedia dump are transformed into plain text and indexed by a Lucene instance. The currently edited document is also converted to plain text and used to search similar articles. If the resulted articles have infoboxes on them - a fact provided by DBpedia -, the hypothesis is that they will be applicable to this document as well. We have tried machine learning techniques to recommend infoboxes and categories but the results were unsatisfactory and we also had to face serious technical problems – concerning mainly memory usage and speed – with a corpus this large. We could not conduct strict numeric measurements on the applicability of the infoboxes recommended by Lucene but user feedback indicates that in certain topics like settlements and biographies it works quite well – that is, the proper infobox is mostly in the top 3-5 recommendations. The recommendation of infrequent infoboxes is less robust in general, but many times it provides with infoboxes previously unknown to the user which they usually consider as an added value.

- Book Recommendation is also implemented by a search to external services, in this case we use RDF data from the British National Bibliography and structured data from the Library of Congress. Both data sets provide a subject field for their entries, in which the important keywords are recorded. The system conducts a search in this field for the most important words and 2-word structures - information provided by the tf-idf and co-occurrence modules.

- PageLink recommendation via simple frequency measures and DBpedia Spotlight. This function is partly based on the tf-idf [6] statistical relevance measure, which is in turn based on the whole Wiki corpus - what we have gathered from the Wiki XML dump - and calculated on-the-fly for the edited article. The tf-idf measure is supplemented by a word co-occurrance measure which is derived from the Wortschatz schema [5]. However, these measures only allow for single and two-word links and are not aware of the language context. A more sophisticated way of offering links is the usage of DBpedia Spotlight, which relies on a number of name-URI associations extracted from titles, redirects and disambiguates, as well as TF*ICF scoring [4] of the target text to choose between possible disambiguation options. The role of this function is to call the user's attention to the possible connections of the current text to other articles.

- Relationship Suggestion. In Semantic MediaWiki[3] instances, users are able to further qualify links between pages in order to indicate the semantic relationship between the concepts represented by those pages. We use the knowledge in DBpedia in order to make suggestions of relationships that are possibly applicable to the current article and the target article suggested by DBpedia Spotlight. In its simplest incarnation, this function can be enabled by querying the DBpedia Ontology to request all relationships that apply to the

[1] http://developer.yahoo.com/search/boss/

[2] http://lucene.apache.org

[3] http://semantic-mediawiki.org/

types of the article being edited and the link being suggested. Suppose somebody is editing the article for `Berlin`, and a link is being added to `Germany`. The system queries the ontology schema and requests all possible relationships between a `City` and a `Country`. These relationships are suggested to the user, who can choose one or none of the options. We believe that by reducing the amount of work necessary to find and assign suitable relationships between concepts, users will be more willing to contribute semantic relationships that are fundamental for querying Wikipedia as a structured database.

4. SYSTEM ARCHITECTURE

Figure 2 depicts the architecture of our system. The Web front end - in this case Sztakipedia toolbar - is communicating with the Aggregation Server, which collects semantic annotations from various Background Modules. Some of these modules - tokenizer, stemmer, tf-idf calculator, co-occurrence calculator, infobox recommendation - are implemented as UIMA [2] Annotation Engines. Others - Category Recommendation, Book Recommendation, DBpedia Spotlight - are provided by REST-based services.

5. CONCLUSION AND PERSPECTIVES

Wikipedia has been a fundamental source of knowledge for many applications, particularly those aiming to model knowledge in structured formats and perform semantic enhancement of content. In this demo we presented how such applications can be used, in turn, to improve the process of knowledge production in Wikipedia. The tool described is live and accessible for everyone, and is already used by a small number of Wiki Editors. It can be installed in 2 minutes, directly from your Wikipedia user page (see video[4]). In future work we plan to perform live re-calculation of the tf-idf and co-occurrence measures; develop the interface of the toolbar into a more proactive recommender; resolve the language dependencies of the tool; and incorporate other data sources.

6. ACKNOWLEDGEMENTS

Héder wishes to thank MTA Sztaki the funding of Sztakipedia project in 2011. Héder's research was also supported by the grant TÁMOP - 4.2.2.B-10/1–2010-0009 (and OTKA K72598). Work on DBpedia Spotlight was supported by the European Commission through the project *LOD2 – Creating Knowledge out of Linked Data*.

7. REFERENCES

[1] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7:154–165, September 2009.

[2] D. Ferrucci and A. Lally. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348, 2004.

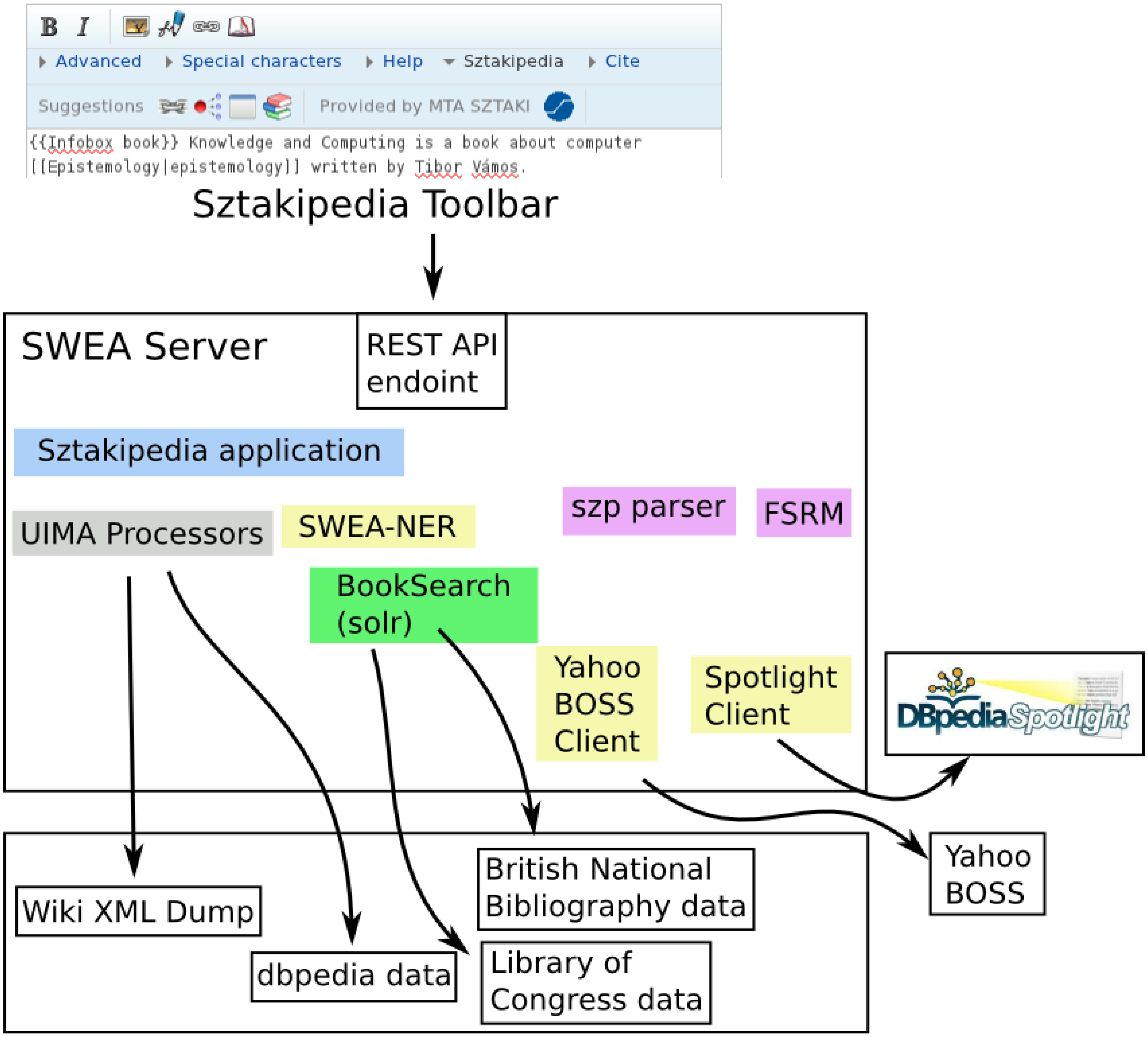
[3] M. Héder. Integrating artificial intelligence solutions into interfaces of online knowledge production. *ICIC Express Letters*, 5(12):4395–4401, 2011.

[4] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*, 2011.

[5] U. Quasthoff, M. Richter, and C. Biemann. Corpus portal for search in monolingual corpora. In *Proceedings of the LREC*, 2006.

[6] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval* 1. *Information processing & management*, 24(5):513–523, 1988.

[4] Video with installation instructions available at: `http://pedia.sztaki.hu/?page_id=9`



{{Infobox book}} Knowledge and Computing is a book about computer
[[Epistemology|epistemology]] written by Tibor Vámos.

Figure 2: The Architecture of Sztakipedia