The Wikipedia Bitaxonomy Explorer

Tiziano Flati and Roberto Navigli

Dipartimento di Informatica
Sapienza Università di Roma

Abstract. We present WiBi Explorer, a new Web application developed in our laboratory for visualizing and exploring the bitaxonomy of Wikipedia, that is, a taxonomy over Wikipedia articles aligned to a taxonomy over Wikipedia categories. The application also enables users to explore and convert the taxonomic information into RDF format. The system is publicly accessible at `wibitaxonomy.org` and all the data is freely downloadable and released under a CC BY-NC-SA 3.0 license.

1 Introduction

Knowledge modeling is a long-standing problem which has been addressed in a variety of ways (see [8] for a survey). If we leave aside knowledge-lean taxonomy learning approaches [9], a typical and widespread model consists of knowledge resources and multilingual dictionaries which provide concepts and relationships between concepts. The scenario is characterized by two types of resources: those, such as BabelNet [6], which provide general untyped relationships, and those, such as DBpedia [1], in which edges model arbitrarily labelled predicates over concepts (e.g., *dbpedia-owl:birthPlace*).

In neither of these resource types, however, is any explicit attention paid to hypernymy as a distinct relation type. Instead, hypernymy has been proven to be a relevant relation type capable of ameliorating systems in several hard tasks in Natural Language Processing [2, 7]. Indeed, even restricting to Wikipedia, no high-quality, large-scale taxonomy is yet available, which exhibits high coverage for both Wikipedia pages and categories.

WiBi [4] is a project set up with the specific aim of providing hypernymy relations over Wikipedia and our tests confirm it as the best current resource for taxonomizing both Wikipedia pages and categories in a joint fashion with state-of-the-art results. Here we present a Web application for visualizing and exploring our bitaxonomy of Wikipedia. The interface also offers a customization of the "view" and allows the export of data into RDF, in line with today's Semantic Web trend.

2 The Wikipedia Bitaxonomy

WiBi [4] is an approach which aims at building a bitaxonomy of Wikipedia, that is, automatically extracting two taxonomies, one for Wikipedia pages and one for Wikipedia categories, aligned to one another.

The bitaxonomy is built thanks to a three-phase approach that i) first builds a taxonomy for the Wikipedia pages, then ii) leverages this partial information to iteratively infer new hypernymy relations over Wikipedia categories while at the same time increasing the page taxonomy, and finally iii) refines the obtained category taxonomy by means of three ad-hoc heuristics that cope with structural problems affecting some categories. As a result, a bitaxonomy is obtained where each element - either page or category - is associated with one or more hypernyms and where elements of one taxonomy are aligned (i.e, linked) to elements of the other taxonomy. In order to transfer hypernymy knowledge from either one of the two Wikipedia sides to the other side, the whole process remarkably, and as a key feature, exploits categorization edges (here called *cross-edges*) manually provided by Wikipedians, which connect any page on one side to its categories on the other side and vice versa. Extensive comparison has been carried out on two datasets of 1,000 pages and categories each, against all the available knowledge resources, including MENTA, DBpedia, YAGO, WikiTaxonomy and WikiNet (for an extensive survey, see [5]). Results show that WiBi surpasses all competitors not only in terms of quality, with the highest precision and recall, but also in terms of coverage and specificity.

3 The demo interface

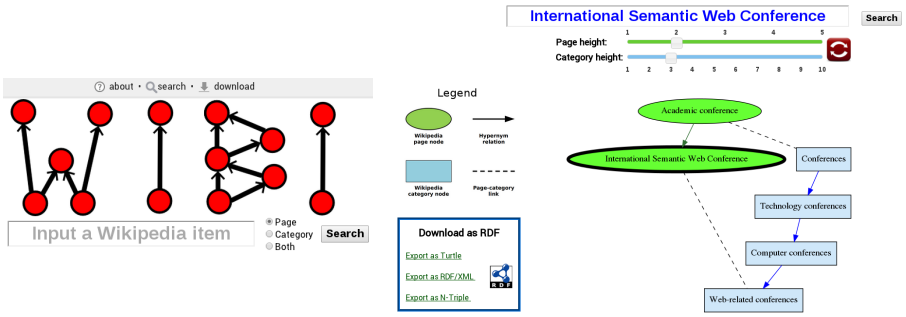
Here we present a Web-based visual explorer for displaying the two aligned taxonomies of WiBi, centered on any given Wikipedia item of interest chosen by the user. The interface easily integrates search facilities with customization tools which personalize the experience from a user's point of view.

The home page. An excerpt of the interface's home page is shown in Fig. 1(a). As can be seen, this page has been kept very clean with as few elements as possible. On the top of the page a navigation bar contains links to i) the *about* page, which contains release information about the website content, ii) a *download* area, where it is possible to obtain the data underlying the interface and iii) the *search* page, which represents the core contribution of this work.

The search page mainly contains a text area in which the user is requested to input her query of interest, additionally opting for searching through either the page inventory, the category inventory or both, thanks to dedicated radio buttons. After the query is sent, the search engine tries to match the input text against the whole database of Wikipedia pages (or categories) and, if a match is found, the engine displays the final result to the user. Otherwise, the query is interpreted as a lemma and the user is returned with the (possible) list of all Wikipedia pages/categories whose lemma matches against the query.

The result page. Starting from the Wikipedia element provided by the user, the objective of the result page is to show a relevant excerpt of the bitaxonomy, that is, the nearest (or more relevant) nodes connected to it, drawn from both of the two taxonomies. To do this, WiBi Explorer performs a series of steps:

1. Start a DFS of maximum length δ_1 from the given element p of a taxonomy. As a result, a subgraph $ST_1 = (SV_1, SE_1)$ is obtained;



(a) WiBi Explorer's home page. (b) Result for the ISWC Wikipedia page.

Fig. 1. The Wikipedia Bitaxonomy Explorer overview.

2. Collect all the nodes $\pi(p)$ belonging to the other taxonomy (i.e, those whose cross-edges are incident to p). Start a DFS of maximum length δ_2 from each element in $\pi(p)$. As a result, a subgraph $ST_2 = (SV_2, SE_2)$ is obtained;
3. Display ST_1 and ST_2, as well as all the possible cross-edges linking nodes of the two subgraphs. Prune out low-connected nodes from the displayed bitaxonomy.

As a result, the interface displays a meaningful excerpt of the two taxonomies, centered on the issued query. The result for the Wikipedia page INTERNATIONAL SEMANTIC WEB CONFERENCE is shown in Fig. 1(b).

Customization of the view Since a user might be interested in a more general view of the bitaxonomy, two additional sliders are provided to the user in order to manually adjust the two maximum depths δ_1 and δ_2 (see Fig. 1(b) on top). Moreover, the interface provides the user with the capability to click on nodes and interactively explore different parts of the taxonomy. The application thus acts as a dynamic explorer that enables users to navigate through the structure of the bitaxonomy and discover new relations as the visit proceeds.

4 Converting data to RDF

Interestingly, data can also be exported in RDF format, in line with recent work on (linguistic) linked open data and the Semantic Web [3]. To this end, the explorer is backed by the Apache Jena framework (`https://jena.apache.org/`) and thus also integrates a single-click functionality that seamlessly converts the displayed data into RDF format. The user can opt for Turtle, RDF/XML or N-Triple format (see blue box in Fig. 1(b), bottom left). An excerpt of a view of the bitaxonomy converted into RDF for the query ISWC is shown in Fig. 2. As can be seen, several namespaces have been used: WiBi specific entities encode Wikipedia items, while standard SKOS's subsumption relations (*skos:narrower* and *skos:broader*) encode is-a relations.

5 Conclusions

We have proposed the Wikipedia Bitaxonomy Explorer, a new, flexible and extensible Web interface that allows the navigation of the recently created Wikipedia

```
@prefix wibi: <http://wibitaxonomy.org/> .
@prefix wibi-model: <http://wibitaxonomy.org/model/wibi#> .
@prefix skos:  <http://www.w3.org/2004/02/skos/core#> .

wibi:International_Semantic_Web_Conference a skos:Concept;
        wibi-model:hasWikipediaCategory  <http://wibitaxonomy.org/Category:Web-related_conferences> ;
        skos:broader               wibi:Academic_conference .

<http://wibitaxonomy.org/Category:Conferences> a skos:Concept;
        wibi-model:hasWikipediaPage  wibi:Academic_conference ;
        skos:narrower      <http://wibitaxonomy.org/Category:Technology_conferences> .
```

Fig. 2. RDF excerpt of the taxonomy view for the ISWC Wikipedia page.

Bitaxonomy [4]. In addition to default settings, several parameters concerning the general appearance of the results can also be customized according to the user's preferences. The demo is available at `wibitaxonomy.org`, it is seamlessly integrated into the BabelNet interface (`http://babelnet.org/`) and the data is freely downloadable under a CC BY-NC-SA 3.0 license.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234.
The authors also acknowledge support from the LIDER project (No. 610782), a Coordination and Support Action funded by the EC under FP7.

References

1. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - a crystallization point for the Web of Data. Web Semantics 7(3), 154–165 (2009)
2. Cui, H., Kan, M.Y., Chua, T.S.: Soft Pattern Matching Models for Definitional Question Answering. ACM Transactions on Information Systems 25(2) (2007)
3. Ehrmann, M., Cecconi, F., Vannella, D., Mccrae, J.P., Cimiano, P., Navigli, R.: Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. In: Proc. of LREC 2014. pp. 401–408. Reykjavik, Iceland
4. Flati, T., Vannella, D., Pasini, T., Navigli, R.: Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. In: Proc. of ACL 2014. pp. 945–955. Baltimore, Maryland
5. Hovy, E.H., Navigli, R., Ponzetto, S.P.: Collaboratively built semi-structured content and Artificial Intelligence: The story so far. Artificial Intelligence 194, 2–27 (2013)
6. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193, 217–250 (2012)
7. Snow, R., Jurafsky, D., Ng, A.: Semantic taxonomy induction from heterogeneous evidence. In: Proc. of the COLING-ACL 2006. pp. 801–808
8. Van Harmelen, F., Lifschitz, V., Porter, B.: Handbook of knowledge representation, vol. 1. Elsevier (2008)
9. Velardi, P., Faralli, S., Navigli, R.: OntoLearn Reloaded: A graph-based algorithm for taxonomy induction. Computational Linguistics 39(3), 665–707 (2013)