A Web Application to Search a Large Repository of Taxonomic Relations from the Web

Stefano Faralli[1], Christian Bizer[1], Kai Eckert[2], Robert Meusel[1], Simone Paolo Ponzetto[1]

[1] University of Mannheim, Germany
{stefano,chris,robert,simone}@informatik.uni-mannheim.de,
[2] Stuttgart Media University, Germany
eckert@hdm-stuttgart.de

Abstract. Taxonomic relations (also known as *isa* or hypernymy relations) represent one of the key building blocks of knowledge bases and foundational ontologies and provide a fundamental piece of information for many text understanding applications. Despite the availability of very large knowledge bases, however, some Natural Language Processing and Semantic Web applications (e.g., Ontology Learning) still require automatic *isa* relation harvesting techniques to cope with the coverage of domain-specific and long-tail terms. In this paper, we present a web application to directly query a very large repository of *isa* relations automatically extracted from the Common Crawl (the largest publicly available crawl of the Web). Our resource can be also downloaded for research purposes and accessed programmatically (we additionally release a Java application programming interface for this purpose).

Keywords: Hearst patterns, hypernym extraction, information extraction and Natural Language Processing techniques for the Semantic Web

1 Introduction

Taxonomic relations play an important role when interpreting data and text that are not already semantically annotated. In fact, to infer the types of entities (be it named entities in text, or entities in semi-structured data) represents a crucial step to understanding the data. Paulheim et al. [5] have shown that adding precise types of instances can lead to a significantly improved performance in many data mining tasks. Moreover, when performing data integration – for instance, of a large collection of tabular datasets into a knowledge base – understanding whether the entities in a table are, for example, cities, states, or mountains, is a very important step towards a high-quality result [7].

While there are quite a few named entity recognition and disambiguation tools that do serve that purpose and exploit knowledge resources such as Wikipedia, DBpedia, or Freebase, a common problem is dealing with the long tail of entities that are not contained in such knowledge bases. These, in fact, have no problems in covering, for instance, major cities ("New York is a city") or celebrities ("Madonna is a singer"), but show limitations with respect to small villages and less known people. Moreover, many common benchmarks for entity linking are also tailored towards popular entities [2]. However, the potential of real-world semantic applications can only be unlocked if

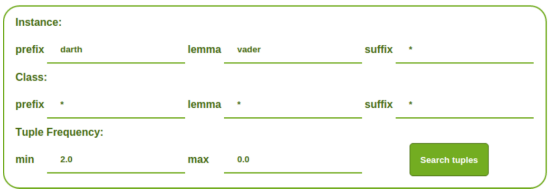

Fig. 1: The main user interface used to query our WebIsA database – for instance, all definitions of the term "darth vader" occurring at least twice.


Found 1754 matches on WebIsADatabase:

	PreTerm	Term	PostTerm	PrecClass	Class	PostClass	Frequency
1	darth	vader		star wars	character		167
2	darth	vader			character		83
3	darth	vader			villain		43
4	darth	vader			none		41
5	darth	vader		iconic	character		34
6	darth	vader			dad		34
7	darth	vader			dad	like any otherexcept	29
8	darth	vader			great		21
9	darth	vader		good	father		21
10	darth	vader		villains playable	character		21
11	darth	vader		bad	guy		20
12	darth	vader		unwanted	change		18
13	darth	vader		other star wars	character		17
14	darth	vader			character	in the star	16
15	darth	vader		original	trilogy		16
16	darth	vader		amazing project	manager		15
17	darth	vader		favorite	character		13
18	darth	vader		lego star wars	universe		13
19	darth	vader		beloved	character		12
20	darth	vader			fan		12
21	darth	vader			boy	name anakin skywalker	12

Fig. 2: The interface to browse the result tuples retrieved for a user query.

these are capable of dealing with the most prominent entities, as well as the long tail. Hence, the necessity of extending existing knowledge bases with hypernymy relations that also cover the long tail entities.

In this paper, we present a web application to query an open knowledge repository consisting of around 400 million *isa* relations, in the form of tuples, which have been automatically extracted from the Common Crawl[1], the largest publicly available crawl of the Web. Our resource is built by combining traditional Hearst-like lexico-syntactic patterns [3] with filtering, duplicate removal and tuple normalization techniques. These methods are applied on web scale using the extraction framework of the WebData-Commons project[2]. Each tuple comes with a rich set of attributes, such as the set of patterns matching the pair, the pay-level domains on which the patterns were matched, the number of occurrences, etc.. In this paper, we present the web application which lets users easily interact with the knowledge repository. A detailed description of our resource construction method and programmatic access can be instead found in [8].

2 A Web Application for the WebIsA Database

The application is designed as a typical *client-server* web application. The server-side implementation includes our Java API [8] to query a MongoDB[3] server serving the

[1] https://commoncrawl.org
[2] http://webdatacommons.org/framework/
[3] https://www.mongodb.com

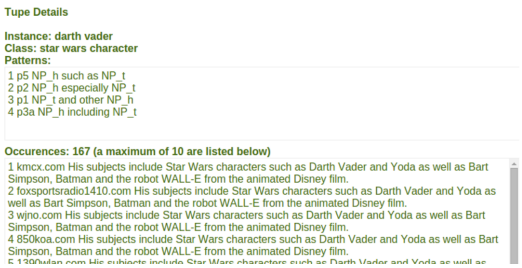

Fig. 3: The interface to browse the additional meta-data of a selected tuple – e.g., "darth vader" *isa* "star wars character".

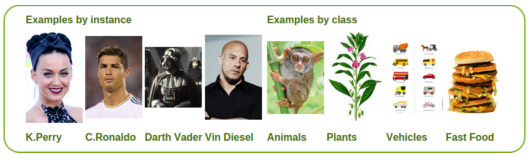

Fig. 4: The set of pre-compiled query examples.

access to an instance of our repository. On the client side, the user is guided by a form-based web page for formulating queries (Figure 1). After submitting a query, the results can be browsed in a tabular format (Figure 2). For each triple in the set of results the table provides the syntactic decomposition of the two noun phrases involved in the *isa* relations into pre-modifiers, head and post-modifiers [6], as well as the frequency of occurrence of the relation in the Common Crawl. The user can also access a detailed view with additional meta-data, namely the patterns matching the relation, the textual contexts of the matching and the pay-level domains indicating the provenance in the corpus (Figure 3). Finally, in order to showcase our application and provide users with some usage examples, we include a few pre-compiled queries – i.e., relations involving instance like "Katy Perry" or "Darth Vader", as well as classes like "Animals", "Plants", and so on (Figure 4).

Note that our system is able to retrieve dozens of tuples also for less popular concepts. As an example, consider the small Italian town of San Sisto (which can be found near to the more famous one Todi). Such a small medieval town, in fact, is not even found within large knowledge bases like Wikipedia (and accordingly YAGO or DBpedia). However, thanks to our knowledge base, we are able to provide the user with useful definitional information nuggets such as the fact that "San Sisto" *isa* "beautiful borgo" – from the pay-level domain `usfreeads.com`, as extracted from the sentence "Sisto is a beautiful borgo very close to the fascinating medieval town of Todi.".

3 Conclusions

In this paper, we have presented a web application to directly query a publicly available knowledge repository containing millions of *isa* relations automatically extracted from

the Common Crawl. Our resource is, to the best of our knowledge, the largest collection of hypernymy relations created from textual resources: accordingly, our web application is meant to provide an easy-to-use user interface for rapid exploration and facilitated access to this very large knowledge resource.

We make our resource freely available to share the wealth of knowledge contained therein, as well as to foster the development of novel knowledge-rich applications that work with the largest and richest textual resource of our time, namely the Web. We believe that our WebIsA database represents a first step towards more complex semantic resources such as web-scale full-fledged taxonomies. In fact, our resource was already successfully used as part of a SemEval competition on taxonomy induction [1], and helped us achieve a competitive performance [4] on challenging benchmarks.

Online Web Application and Downloads

Our web interface is available at `http://webisadb.webdatacommons.org/`. All the resources described in this paper are freely available under a CC BY-NC-SA 3.0 license at `http://webdatacommons.org/isadb/`, where we additionally provide a Java application programming interface for programmatic access from client applications, as well as the source code of the extraction framework.

Acknowledgements

This work was partially funded by the Junior-professor funding programme of the Ministry of Science, Research and the Arts of the state of Baden-Württemberg, Germany (project “Deep semantic models for high-end NLP applications”). Part of the computational resources were provided by an Amazon AWS in Education Grant award.

References

1. Bordea, G., Lefever, E., Buitelaar, P.: SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In: Proc. SemEval. pp. 1081–1091 (2016)
2. van Erp, M., Mendes, P., Paulheim, H., Ilievski, F., Plu, J., Rizzo, G., Waitelonis, J.: Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In: Proc. LREC (2016)
3. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proc. COLING. pp. 539–545 (1992)
4. Panchenko, A., Faralli, S., Ruppert, E., Remus, S., Naets, H., Fairon, C., Ponzetto, S.P., Biemann, C.: TAXI at SemEval-2016 Task 13: A taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In: Proc. SemEval. pp. 1320–1327 (2016)
5. Paulheim, H., Fürnkranz, J.: Unsupervised Generation of Data Mining Features from Linked Open Data. In: Proc. WIMS. pp. 31:1–31:12 (2012)
6. Ponzetto, S.P., Strube, M.: Taxonomy induction based on a collaboratively built knowledge repository. ArtInt 175(9-10), 1737–1756 (2011)
7. Ritze, D., Lehmberg, O., Bizer, C.: Matching HTML tables to DBpedia. In: Proc. WIMS. pp. 10:1–10:6 (2015)
8. Seitner, J., Bizer, C., Eckert, K., Faralli, S., Meusel, R., Paulheim, H., Ponzetto, S.P.: A large database of hypernymy relations extracted from the web. In: Proc. LREC (2016)