WebIsALOD: Providing Hypernymy Relations extracted from the Web as Linked Open Data

Sven Hertling and Heiko Paulheim

Data and Web Science Group, University of Mannheim, Germany
{sven,heiko}@informatik.uni-mannheim.de

Abstract. Hypernymy relations are an important asset in many applications, and a central ingredient to Semantic Web ontologies. The IsA database is a large collection of such hypernymy relations extracted from the Common Crawl. In this paper, we introduce WebIsALOD, a Linked Open Data release of the IsA database, containing 400M hypernymy relations, each provided with rich provenance information. As the original dataset contained more than 80% wrong, noisy extractions, we run a machine learning algorithm to assign confidence scores to the individual statements. Furthermore, 2.5M links to DBpedia and 23.7k links to the YAGO class hierarchy were created at a precision of 97%. In total, the dataset contains 5.4B triples.

Keywords: Hypernyms, Hearst Patterns, Linked Dataset

1 Introduction

Hypernymy relations are an important asset in many applications, and a central ingredient to Semantic Web ontologies. They can be used in various applications – for example in named entity recognition and disambiguation tools, or as background knowledge to improve the performance of data mining tasks [9]. Often the approaches rely on knowledge bases like Wikidata, DBpedia or YAGO, which are good at head entities, but lack coverage and level of detail for tail entities. While hypernymy datasets have been created, such as LHD [6], they rely on entities which are contained as instances in Wikipedia, and hence expose the same bias towards head entities [10]. To fill that gap, Seitner et al. created a large database of hypernymy relations extracted from the Web [12], the IsADB.

The main idea of the IsADB is to extract hypernymy relations from a huge and fixed web crawl called CommonCrawl[1]. The extraction method is based on 58 Hearst-like lexico-syntactic patterns which are frequent patterns to describe type relations. For example, the sentence *Still, people use Gmail and other Web services* implies the hypernymy relation between *Gmail* and *Web service*, which can be captured by the pattern *NP and other NP*.[2]

In this work, we present a Linked Data endpoint to the IsADB, following the best practices for Linked Open Data [11]. The dataset provides access via HTTP

[1] https://commoncrawl.org

[2] *NP* stands for *noun phrase.*

URIs and a SPARQL endpoint, and it is interlinked to DBpedia [7] and YAGO [13]. Furthermore, it provides rich provenance information for each hypernymy relation, which capture

- the pre modifier, post modifier, and head noun for both the hypernym and the hyponym. In the example above, *service* is the head noun of the hypernym, and *Web* is its pre-modifier;
- the set of pattern ids (PIDs) matching the hypernymy relation;
- the set of sentences which was used for the extraction;
- the set of pay-level domains (PLDs) on which the sentences appear; and
- the absolute number of hyponym-hypernym pair occurrences (frequency).

That information is also used to apply machine learning for computing confidence scores for all relations. Hence, the provided dataset also allows for quality-based filtering on the provided information.

The rest of this paper is structured as follows. Section 2 analyses the original (non-LOD) IsADB dataset. Section 3 introduces the model used for providing WebIsALOD. Section 4 describes the interlinking to DBpedia and YAGO, and section 5 describes the method applied for computing confidence scores. Section 6 provides a first content profile of the resulting dataset.

2 The Original IsADB Dataset

The original dataset contains 400,533,808 relations, 120,992,255 unique hyponyms, and 107,691,822 unique hypernyms collected with 58 different patterns. To assess the quality of the dataset i.e. how many relations are actually valid, a crowd-sourced survey via Amazon Mechanical Turk (MTurk)[3]was conducted.

The participants of the survey were presented sentences in the form *hyponym is a hypernym*, constructed from random pairs of hyponyms and hypernyms from the database. For each of those sentences, they could answer “Yes”, “Uncertain”, or “No”.[4]

In order to gain stable results, each sentence was rated by nine different workers. The final label (“true”,”false”,”uncertain”) was assigned by majority voting.

For estimating the fraction of correct axioms, 500 randomly sampled hypernymy relations from the original dataset were presented to the participants. Additionally, we estimated the quality of the dataset at different lower thresholds of two key figures: 1) the amount of patterns which are used for the extraction and 2) the amount of pay-level domains. Both can be understood as (weak) indicators for the correctness of a relation. A dataset with the given threshold t is

[3] https://www.mturk.com

[4] We restricted the workers to have a 95% approval rate and a minimum of 100 approved HITs (human intelligence tasks), following the recommendations by [5] and [3], and restricted their location to the US to attract a large fraction of native speakers.

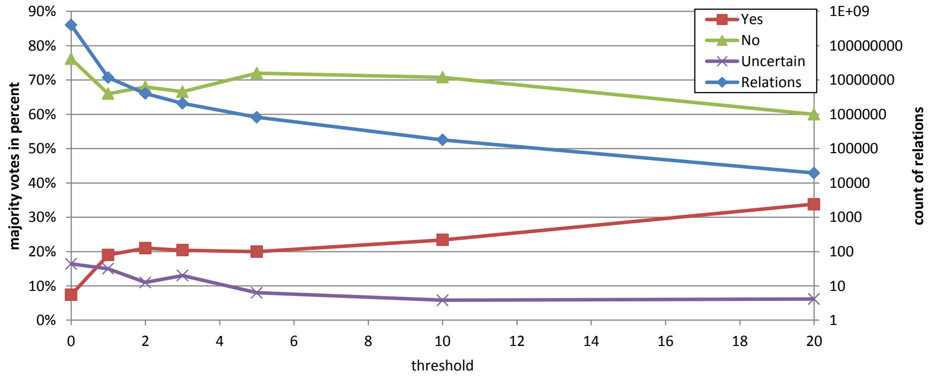

Fig. 1. Quality ratings and overall count of relations for different thresholds of pattern and PLD spread.

defined by

$$dataset(t) = \{r \in R \mid |r.pld| > t \wedge |r.pid| > t\}$$

Seven thresholds were chosen for evaluation: $0, 1, 2, 3, 5, 10, 20$, where 0 corresponds to the full dataset. Figure 1 shows the amount of relations in the corresponding set as well as the percentage of the majority vote. It shows that there is a steep quality increase when stepping from a threshold 0 to a threshold of 1, while there is only moderate gain for lower thresholds of 10 and 20. On the other hand, increasing the threshold drastically decreases the number of relations from 400M to little more than 10k. When utilizing all the data (i.e., imposing a threshold of 0), 7.4% of 400,533,808 relations are correct. Extrapolating these values results in 29,639,501 true relations.

These figures show simply applying a lower threshold on $|r.pld|$ and $|r.pid|$ would remove a large amount of noise, it also reduces the dataset size drastically, limiting its utility. Hence, we chose a different approach: we train a machine learning model to assign confidence scores to the relations. This model is applied to the full dataset, allowing users of the dataset to impose a quality threshold and trade off coverage and accuracy indivdually given their task at hand.

3 Dataset Modeling and Provision

A major goal of providing the WebIsALOD dataset is to not only provide access to the hypernymy relations as such, but also to rich metadata for those relations.

Fig. 2 shows an excerpt of the hypernymy relation used above (*Web service* is a hypernym of *GMail*), together with a subset of its metadata.

For modeling the actual hypernymy relations, we inspected several alternatives, including `rdf:type`, `rdfs:subClassOf`, and `skos:broader`. Since we have both instances (like *Gmail*) as well as classes (like *Web service*) in our dataset,

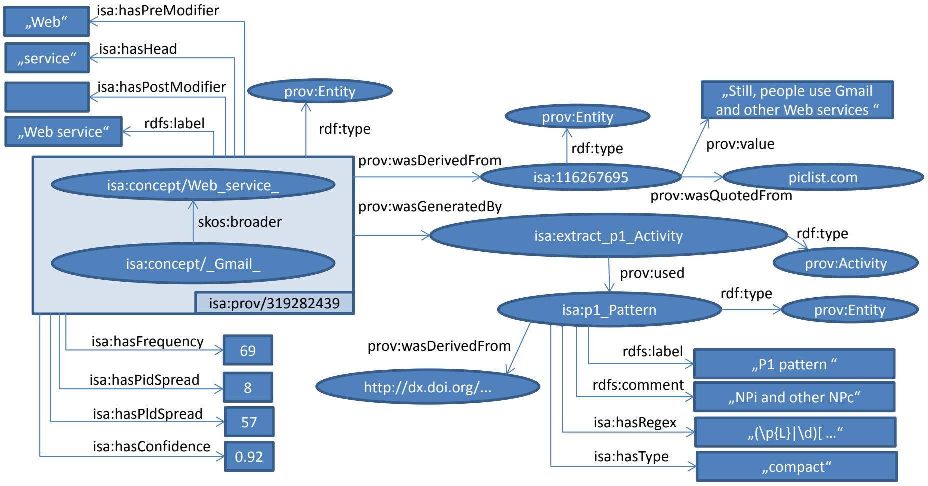

Fig. 2. Example depiction of a hypernymy relation with its metadata

and we cannot trivially distinguish them, using `skos:broader` has been considered as the most appropriate relation.

Each hypernymy relation is stored in its own named graph [2], indicated by the rectangular box in Fig. 2. For each hyponym and hypernym, we provide the head, premodifier, and postmodifier, together with the actual label. Hyponyms and hypernyms are linked to DBpedia instances and YAGO classes, as described in section 4. Statements about the provenance of the hypernymy relation are then made about that named graph, including

- the originating sentence from which the relation was extracted, and the Web page on which the sentence was found,
- the pattern which was used to extract the relation, together with a description, a regular expression formulation, and a link to the literature source in which the pattern was proposed, and
- statistical metadata, such as the global frequency, the PID and PLD spread ($|r.pid|$ and $|r.pld|$), and the confidence score computed (see section 5).

The data is provided as Linked Open Data, using dereferencable URIs, as a dump for download, as well as through a SPARQL endpoint.[5] The latter also allows the user for filtering by a specific confidence threshold in order to control the quality of the returned information, and trade off coverage against precision per use case. The source code as well as the templates and results of the crowdsourced survey is available at github.[6]

[5] http://webisa.webdatacommons.org/
[6] https://github.com/sven-h/webisalod

4 Linking to DBpedia and YAGO

In order to follow the Linked Data best practices and provide interlinks to other datasets, we chose two different datasets as link targets: DBpedia for instances, as it is the de facto interlinking hub of the Linked Open Data cloud [11], and YAGO for classes, as it provides one of the richest general purpose class hierarchies.

For performing the interlinking of instances, three approaches were tried:

- Using plain string matching on lower cased strings
- Using the DBpedia surface forms [1]
- Using DBpedia Spotlight [8] on the original sentences

To evaluate the different strategies, we created another Amazon MTurk survey, where we asked the annotators to provide Wikipedia pages and categories for both the hyponym and the hypernym of a relation. The Wikipedia pages were then translated to DBpedia URIs as a gold standard to test the interlinking of instances. The plain string matching clearly outperformed the other two with an F1 score of 0.97 (precision 0.97, recall 0.97, compared to an F1 score of 0.59 for surface forms and 0.54 for DBpedia Spotlight). Hence, we used this approach to create in total 2,593,181 instance interlinks.

Since the results were satisfying for instance matching, and neither the surface forms approach nor DBpedia Spotlight can produce links to YAGO classes (or Wikipedia categories), we also use this approach to create links to YAGO classes. Since those are derived from Wikipedia categories, we use the MTurk gold standard for evaluation. We achieve an F1 score of 0.72 (precision 0.93, recall 0.59), and create a total of 23,771 links to YAGO classes.

5 Computing Confidence Scores

For computing the confidence scores, we trained a machine learning classifier on the labels ("correct","incorrect","uncertain") assigned in the initial Amazon MTurk evaluation on dataset(1). In total, the dataset contains 95 correct and 330 incorrect instances; the 75 instances with a majority of "uncertain" or an equal share of "correct" and "incorrect" were discarded. By classifying the relations as correct or incorrect, the classifier's confidence score for the label *correct* can be used as a confidence score for the relation itself.

We used six different classifiers and performed parameter tuning in 10-fold cross validation:

- Decision Trees optimized by minimum leaf size and maximum depth of tree (1-20)
- Gradient Boosted Trees optimized by maximum depth (1,5,9,12,16,20) and number of trees (20,40,60,80,100)
- RandomForest optimized by number of trees (1-100 with 10 steps) and minimum leaf size (1-10)
- Naive Bayes (without specific parameter tuning)

Table 1. Results of different classifiers and feature sets using 10-fold cross validation on the gold standard of dataset(1) (AUC 1), and evaluated on the gold standard of dataset(0) (AUC 0)

ML approach	FS 1		FS 1+2		FS 1+2+3	
	AUC 1	AUC 0	AUC 1	AUC 0	AUC 1	AUC 0
Decision Tree	0.7572	0.6063	0.7801	0.6544	0.7547	0.6742
GBT	0.8032	0.6490	0.8176	0.6783	0.8086	0.6954
RandomForest	0.8287	0.7020	**0.8446**	0.6427	0.8377	**0.7246**
Naive Bayes	0.5782	0.5080	0.5782	0.5080	0.6338	0.5183
SVM	0.8194	0.6444	0.8410	0.6994	0.8411	0.6863
Neural Net	0.7783	0.6080	0.7753	0.6684	0.7757	0.5988

- SVM with Radial Base Function kernel, and C and gamma tuned according to [4]
- Neural Network with one hidden layer in two different sizes $F/2+2$,$sqrt(F)$, and two hidden layers of $F/2$ and $sqrt(F)$, where F denotes the number of features

Table 1 lists the results of all machine learning approaches together with three different feature sets:

- FS1 consists of frequency, amount of patterns, amount of pay-level domains for the relation itself as well as for the relation without pre and post modifier, and a binary value for each pattern indicating if it is extracting the relation or not.
- FS2 adds features derived from the hypernym and hyponym itself, i.e., amount of tokens, average token length, and the existence of a pre and a post modifier.
- FS3 adds features derived from the originating sentences, in particular the token distance between the hyponym and the hypernym. We use the minimum, maximum, and average across all sentences, as well as the number of sentences a pattern spans.

We used the gold standard crowd sources for dataset(1) (i.e., the dataset with a minimum threshold of 1 for $|r.pid|$ and $|r.pld|$) for training, and tested it both in cross validation and on the gold standard for dataset(0) (i.e., the full dataset). Table 1 shows the results for the area under the ROC curve (AUC). We chose an optimization towards ROC AUC because this is an indicator of the quality of confidence scores, and hence the selection criterion for a classification algorithm. Based on those results, we chose the RandomForest classifier utilizing the full set of features trained on the gold standard for dataset(1) to create confidence scores for the full dataset.[7]

[7] The reason why we did not use the gold standard of the full dataset for training is its imbalance (cf. section 2), i.e., the number of positive examples (only 37 out of 500) is too low for learning a meaningful model.

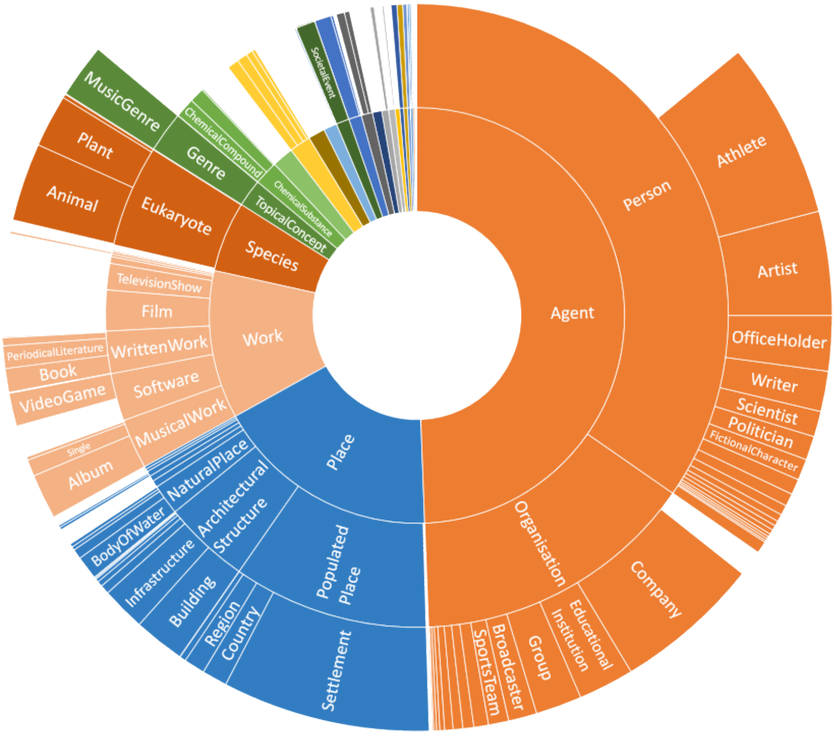

Fig. 3. Type breakdown of the instances linked to DBpedia

6 Analysis of Resulting Dataset

The final resulting dataset consists of 400.5M hypernymy relations, together with a confidence score and metadata, as well as 2,593,181 instance links to DBpedia and 23,771 class links to YAGO. All in all, the dataset consists of 5.4B triples.

In order to obtain a first content profile, we analyzed the fraction of instances which are linked to and typed in DBpedia, and analyzed the type hierarchy in DBpedia to estimate the distribution of those entities. That resulting distribution is depicted in Fig. 3.

We can observe that about half of the information is about persons and organizations. Places, works, and species make up for 18%, 12%, and 5%, respectively, while the rest is a mix of other types.

7 Conclusion and Outlook

In this paper, we have introduced a new dataset of hypernymy relations extracted from the Web, provided as Linked Data with provenance information and interlinks to DBpedia and YAGO.

The dataset has room for improvement in various directions. Examples of ongoing and future work include the learning of better scoring models and the induction of a type hierarchy, where the latter also includes the subtask of automatically distinguishing *subclass of* and *instance of* relations.

Another crucial issue is the identification of homonyms in the dataset. Given the two assertions *Bauhaus is a goth band* and *Bauhaus is a German school*, it is clear that the subjects are two disjoint instances, while *Bauhaus is a goth band* and *Bauhaus is a post-punk band* are not. Identifying such homonyms, e.g., by exploiting upper ontologies, is an ongoing effort.

References

1. Bryl, V., Bizer, C., Paulheim, H.: Gathering alternative surface forms for dbpedia entities. In: NLP-DBPEDIA@ISWC. pp. 13–24 (2015)
2. Carroll, J.J., Bizer, C., Hayes, P., Stickler, P.: Named graphs, provenance and trust. In: International Conference on World Wide Web. pp. 613–622. ACM (2005)
3. Hauser, D.J., Schwarz, N.: Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. Behavior Research Methods 48(1), 400–407 (2016)
4. Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification (2003)
5. Kazai, G.: In Search of Quality in Crowdsourcing for Search Engine Evaluation, pp. 165–176. Springer Berlin Heidelberg (2011)
6. Kliegr, T., Zamazal, O.: Lhd 2.0: A text mining approach to typing entities in knowledge graphs. Web Semantics: Science, Services and Agents on the World Wide Web 39, 47 – 61 (2016)
7. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal 6(2) (2013)
8. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: 7th International Conference on Semantic Systems. pp. 1–8. ACM (2011)
9. Paulheim, H., Fümkranz, J.: Unsupervised generation of data mining features from linked open data. In: 2nd International Conference on Web Intelligence, Mining and Semantics. p. 31. ACM (2012)
10. Ringler, D., Paulheim, H.: One knowledge graph to rule them all? analyzing the differences between dbpedia, yago, wikidata & co. In: 40th German Conference on Artificial Intelligence (2017)
11. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the Linked Data Best Practices in Different Topical Domains. In: International Semantic Web Conference. LNCS, vol. 8796 (2014)
12. Seitner, J., Bizer, C., Eckert, K., Faralli, S., Meusel, R., Paulheim, H., Ponzetto, S.: A large database of hypernymy relations extracted from the web. In: Language Resources and Evaluation Conference, Portoroz, Slovenia (2016)
13. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In: 16th international conference on World Wide Web. pp. 697–706 (2007)