DBpediaNYD – A Silver Standard Benchmark Dataset for Semantic Relatedness in DBpedia

Heiko Paulheim

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

Abstract. Determining the semantic relatedness (i.e., the strength of a relation) of two resources in DBpedia (or other Linked Data sources) is a problem addressed by quite a few approaches in the recent past. However, there are no large-scale benchmark datasets for comparing such approaches, and it is an open problem to determine which of the approaches work better than others. Furthermore, larget-scale datasets for training machine learning based approaches are not available. DBpediaNYD is a large-scale synthetic silver standard benchmark dataset which supports contains symmetric and asymmetric similarity values, obtained using a web search engine.

Keywords: Semantic relatedness, triple ranking, Normalized Google Distance

1 Motivation

When looking at very prominent resources in DBpedia [1], the number of incoming and outgoing properties and, hence, related concepts can be fairly large. For example, the resource `dbpedia:Germany` has a total of 246 outgoing and 64,533 ingoing properties. In total, 36,365 resources are related to `dbpedia:Germany`.[1]

These numbers are fairly large for many practical purposes. For example, for human consumption, displaying more than 60,000 triples about `dbpedia:Germany` is not useful. Thus, mechanisms for assessing the relevance of a certain connected resource to a resource in question are required. Such mechanisms can compute the strength of each related resource and, based on those strengths, select the most relevant ones.

In the recent past, several approaches have been discussed to address this problem, using methods from social network analysis [10], network theory [3], machine learning [6, 9], or external sources, such as web search engines [8] or crowdsourcing via games with a purpose [5].

However, none of those approaches have used the same dataset for validating their results. That makes it impossible to rate those approaches and decide which ones work better than others. In most of the papers, an evaluation is reported

[1] Numbers according to DBpedia 3.8

which involves a number of human experts creating a very limited gold standard based on a few resources only, or, even worse, to only judge the respective tool's results.

Creating an encompassing gold standard manually for measuring relatedness of DBpedia resources is expensive and tedious work. The *DBpediaNYD* dataset[2] is a machine generated silver standard which can be used as a standard benchmark dataset for such approaches. While it is not as perfect as an expert generated dataset, it is an interesting complement for such expert-driven evaluations, since it contains several thousands pairs of results and thus allows for larger-scale evaluation.

Furthermore, most human created datasets are comparably small (e.g., the dataset obtained in [5] through crowdsourcing comprises 183 triples). Datasets used in the NLP community, like the classic datasets by Rubenstein and Goodenough [11] or Miller and Charles [7], never exceed a few hundred pairs of words[3]. In contrast, the DBpediaNYD dataset consists of almost 7,000 pairs of resources, and the method used for its construction scales up to arbitrary sizes. Thus, it can be used in scenarios where larger scale datasets are required, e.g., training of machine learning models.

2 The Dataset

For automatically creating a large-scale benchmark dataset, we exploit web search engines, in particular *Yahoo!*, to compute semantic similarities using the Normalized Google Distance[4], defined in [2] as

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}, \tag{1}$$

In a nutshell, $NGD(x,y)$ measures the likelihood of two terms x and y appearing on the same website. $f(x)$, $f(y)$ and $f(x,y)$ are the number of pages containing x, y, and both x and y. M is the total size of the search engine used for computation.[5] In [4], it has been shown that this formula, in particular using *Yahoo!* as a web search engine, yields similarity values which are reasonably correlated to those defined by human experts.

The standard NGD formula always yields symmetric values. In [12], a directed, hence asymmetric variant of NGD has been proposed, defined as:

$$\overrightarrow{NGD}(x,y) = \frac{\log f(x) - \log f(x,y)}{\log M - \log f(y)} \tag{2}$$

$\overrightarrow{NGD}(x,y)$ measures the likelihood of a web page containing a term x also containing a term y.

[2] `http://wiki.dbpedia.org/FindRelated/files?get=dataset.zip`

[3] See `http://www.semantic-measures-library.org/sml/index.php?q=benchmarks`

[4] Since we use *Yahoo!*, it is actually a Normalized Yahoo Distance, hence the name *DBpediaNYD* for the dataset.

[5] We use the value $10,000,000,000$ for *Yahoo!*, as indicated at `http://www.worldwidewebsize.com/` as of April 15th, 2013.

Asymmetric distances can be useful for the task of rating how relevant a connected resource is for a resource under inspection. For example, *MySQLManager* is a (less well known) application developed by *Apple Inc.* When computing resources relevant for *MySQLManager*, *Apple Inc.* is a relevant concept. However, the list of the most important resources for *Apple Inc.* is more likely expected to contain resources such as *iPhone*, *iPad*, or *Steve Jobs*. Asymmetric similarity distances can cover those cases. Thus, the *DBpediaNYD* dataset contains both standard and directed NGD values.

For creating the DBpediaNYD dataset, we randomly sampled 10,000 statements from DBpedia, using the mapped and raw infobox properties. Using the labels of both the subject and the object, we used the *Yahoo BOSS* search engine[6] to determine the three count values $f(x)$, $f(y)$, and $f(x, y)$ that are used to compute both symmetric and asymmetric NGD.

From those triples, we removed all those with nonsensical results, i.e., those where $f(x, y) > f(x)$ or $f(x, y) > f(y)$ holds.[7] We ended up with a total of 6,942 triples that have useful search engine counts, which leads to a dataset of 6,942 triples for computing NGD, and 13,884 triples for computing asymmetric NGD (since each triple may be used in both directions here).

An example from the dataset is the pair *John Lennon* and *Yoko Ono*. The symmetric distance, as well as the asymmetric distance *John Lennon* $\rightarrow$ *Yoko Ono*, is 0.18, while the asymmetric distance *Yoko Ono* $\rightarrow$ *John Lennon* is only 0.03. This points at the fact that *Yoko Ono*, despite her own artistic career, is most often mentioned on web pages as the wife of *John Lennon*, while *John Lennon* is more often mentioned in other contexts as well (e.g., for his work with the Beatles).

3 Possible Usages

There are two main usages of the DBpediaNYD dataset: benchmarking algorithms and approaches for computing semantic relatedness, and training and evaluating machine learning based approaches.

As indicated in section 1, there are quite a few approaches to determining similarity in DBpedia, which have not been compared to each other with respect to result quality. The DBpediaNYD dataset makes a comparison of those approaches possible.

Nevertheless, the dataset should be used only as a complement to human created gold standards, since it is only a silver standard, and since some of the approaches discussed above propose the use of web search engines themselves [8], which would give them an unfair advantage in a comparative evaluation.

While running our approach on the whole of DBpedia is a tempting vision, it is not feasible with current search engines. Most search engine providers charge

[6] `http://developer.yahoo.com/boss/search/`

[7] For a discussion of that problem, see `http://searchengineland.com/why-google-cant-count-results-properly-53559`

Table 1. Top 5 results from *DBpedia FindRelated* for three example resources. Conway Berners-Lee and Mary Lee Woods are Tim Berners-Lee's parents.

	Apple, Inc.	*Semantic Web*	*Tim Berners-Lee*
	Symmetric	*Symmetric*	*Symmetric*
1	Apple Wireless Keyboard	Turtle (syntax)	WWW Consortium
2	MacBook Pro	N-Triples	Conway Berners-Lee
3	MobileMe	RDF Schema	Mary Lee Woods
4	iCloud	RDFa	WWW Foundation
5	Apple A5	SOA4all	Weaving the Web (book)
	Asymmetric	*Asymmetric*	*Asymmetric*
1	IOS SDK	Ontotext	Line Mode Browser
2	IMac G3	TriX (syntax)	libwww
3	IPod Touch	Notation3	Unitarian Universalism
4	IPod	TriG (syntax)	Computer Scientist
5	MobileMe	International Semantic Web Conference	WWW Consortium

money for programmatic search engine requests (e.g., Yahoo BOSS! charges $0.80 for 1,000 requests), which makes the computation on all of DBpedia very costly.[8]

Machine-learning based approaches for computing semantic relatedness require larger-scale training datasets, which are expensive to obtain through human experts. The DBpediaNYD dataset may be used as such a training set, and it is large enough to reserve certain portions of the dataset for cross validation. One example is the *DBpediaFindRelated* service,[9] which uses the DBpediaNYD dataset for training a support vector machine model, and serves a list of resources related to a given resource, ranked by the computed similarity (either symmetric or asymmetric). The resulting model provides a moderate positive correlation with the original dataset and is thus good enough for many practical use cases, e.g., for ordering triples in a user interface. Table 1 shows the top 5 results of the service for three example resources, both with the symmetric and asymmetric model.

However, approaches using the DBpediaNYD dataset itself as a training set should not be included in benchmarks using the same dataset as an evaluation dataset, due to overfitting. In that case, the dataset should be split into a training and an evaluation portion, or a separate, blind evaluation dataset should be obtained using the same method, but a different sample of resources.

4 Conclusion and Outlook

In this paper, we have introduced the DBpediaNYD dataset, a machine-generated benchmark silver standard dataset which contains symmetric and asymmetric

[8] To the best of our knowledge, there are no cost-free alternatives. Free search engines, such as FAROO, do not provide search result counts and can thus not be used to compute NGD.

[9] `http://dbpedia.org/FindRelated`

distances for 6,942 random pairs of DBpedia resources. The dataset may be used for benchmarking approaches to compute semantic relatedness in DBpedia, and as a training dataset for machine learning based approaches.

Although the correlation of Normalized Google Distance computed with Yahoo! and human ratings has already been shown, e.g., in [4], we aim at validating our approach using smaller, manually created gold standards. Furthermore, it would be beneficial to compare and use the results of several search engines, both for assessing the robustness of the dataset, as well as for using averages from different search engines.

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*, 7(3):154–165, 2009.
2. Rudi Cilibrasi and Paul M. B. Vitányi. The google similarity distance. *CoRR*, abs/cs/0412098, 2004.
3. Thomas Franz, Antje Schultz, Sergej Sizov, and Steffen Staab. Triplerank: Ranking semantic web data by tensor decomposition. In *Proceedings of the 8th International Semantic Web Conference (ISWC 2009)*, Lecture Notes in Computer Science, pages 213–228. Springer, 2009.
4. Jorge Gracia and Eduardo Mena. Web-based measure of semantic relatedness. In *Web Information Systems Engineering-WISE 2008*, pages 136–150. Springer, 2008.
5. Jörn Hees, Thomas Roth-Berghofer, Ralf Biedert, Benjamin Adrian, and Andreas Dengel. Betterrelations: Collecting association strengths for linked data triples with a game. In *Search Computing*, volume 7538 of *Lecture Notes in Computer Science*, pages 223–239, 2012.
6. Edgar Meij, Marc Bron, Laura Hollink, Bouke Huurnink, and Maarten Rijke. Learning semantic query suggestions. In *Proceedings of the 8th International Semantic Web Conference (ISWC 2009)*, Lecture Notes in Computer Science, pages 424–440. Springer, 2009.
7. George A Miller and Walter G Charles. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28, 1991.
8. Roberto Mirizzi, Azzurra Ragone, Tommaso Di Noia, and Eugenio Di Sciascio. Ranking the linked data: the case of dbpedia. In *International Conference on Web Engineering (ICWE 2010)*, Lecture Notes in Computer Science, pages 337–354. Springer, 2010.
9. Kunal Mulay and P. Sreenivasa Kumar. Spring: Ranking the results of sparql queries on linked data. In *The 17th International Conference on Management of Data (COMAD 2011)*, 2011.
10. Bernardo Pereira Nunes, Ricardo Kawase, Stefan Dietze, Davide Taibi, Marco Antonio Casanova, and Wolfgang Nejdl. Can entities be friends? In *Proceedings of the Web of Linked Entities Workshop in conjuction with the 11th International Semantic Web Conference*, volume 906 of *CEUR-WS.org*, pages 45–57, 2012.
11. Herbert Rubenstein and John B Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.
12. Wei Lee Woon and Stuart Madnick. Asymmetric information distances for automated taxonomy construction. *Knowledge and information systems*, 21(1):91–111, 2009.