DBpedia: A Multilingual Cross-Domain Knowledge Base

Pablo N. Mendes[1], **Max Jakob**[2], **Christian Bizer**[1]

[1] Web Based Systems Group, Freie Universität Berlin, Germany
[2] Neofonie GmbH, Berlin, Germany
first.last@fu-berlin.de, first.last@neofonie.de

Abstract

The DBpedia project extracts structured information from Wikipedia editions in 97 different languages and combines this information into a large multi-lingual knowledge base covering many specific domains and general world knowledge. The knowledge base contains textual descriptions (titles and abstracts) of concepts in up to 97 languages. It also contains structured knowledge that has been extracted from the infobox systems of Wikipedias in 15 different languages and is mapped onto a single consistent ontology by a community effort. The knowledge base can be queried using the SPARQL query language and all its data sets are freely available for download. In this paper, we describe the general DBpedia knowledge base and as well as the DBpedia data sets that specifically aim at supporting computational linguistics tasks. These task include Entity Linking, Word Sense Disambiguation, Question Answering, Slot Filling and Relationship Extraction. These use cases are outlined, pointing at added value that the structured data of DBpedia provides.

Keywords: Knowledge Base, Semantic Web, Ontology

1. Introduction

Wikipedia has grown into one of the central knowledge sources of mankind and is maintained by thousands of contributors. Wikipedia articles consist mostly of natural language text, but also contain different types of structured information, such as infobox templates, categorization information, images, geo-coordinates, and links to external Web pages. The DBpedia project (Bizer et al., 2009) extracts various kinds of structured information from Wikipedia editions in multiple languages through an open source extraction framework. It combines all this information into a multilingual multidomain knowledge base. For every page in Wikipedia, a Uniform Resource Identifier (URI) is created in DBpedia to identify an entity or concept being described by the corresponding Wikipedia page. During the extraction process, structured information from the wiki such as infobox fields, categories and page links are extracted as RDF triples and are added to the knowledge base as properties of the corresponding URI.

In order to homogenize the description of information in the knowledge base, a community effort has been initiated to develop an ontology schema and mappings from Wikipedia infobox properties to this ontology. This significantly increases the quality of the raw Wikipedia infobox data by typing resources, merging name variations and assigning specific datatypes to the values. As of March 2012, there are mapping communities for 23 languages [1]. The English Language Wikipedia, as well as the Greek, Polish, Portuguese and Spanish language editions have mapped (to the DBpedia Ontology) templates covering approximately 80% of template occurrences [2]. Other languages such as Catalan, Slovenian, German, Georgian and Hungarian have covered nearly 60% of template occurrences. As a consequence, most of the facts displayed in Wikipedia pages via templates are being extracted and mapped to a unified schema.

In this paper, we describe the DBpedia knowledge base and the DBpedia data sets that specifically aim at supporting computational linguistics tasks. These include the Lexicalization, Topic Signatures, Topical Concepts and Grammatical Gender data sets.

2. Resources

2.1. The DBpedia Ontology

The DBpedia Ontology organizes the knowledge on Wikipedia in 320 classes which form a subsumption hierarchy and are described by 1,650 different properties. It features labels and abstracts for 3.64 million things in up to 97 different languages of which 1.83 million are classified in a consistent ontology, including 416,000 persons, 526,000 places, 106,000 music albums, 60,000 films, 17,500 video games, 169,000 organizations, 183,000 species and 5,400 diseases. Additionally, there are 6,300,000 links to external web pages, 2,724,000 links to images, 740,000 Wikipedia categories and 690,000 geographic coordinates for places.

The alignment between Wikipedia infoboxes and the ontology is done via community-provided mappings that help to normalize name variation in properties and classes. Heterogeneities in the Wikipedia infobox system, like using different infoboxes for the same type of entity (class) or using different property names for the same property, can be alleviated in this way. For example, '*date of birth*' and '*birth date*' are both mapped to the same property `birthDate`, and infoboxes '*Infobox Person*' and '*Infobox FoundingPerson*' have been mapped by the DBpedia community to the class `Person`. DBpedia Mappings currently exist for 23 languages, which means that other infobox properties such as '*data de nascimento*' or '*Geburtstag*' – date of birth in Portuguese and German, respectively – also get mapped to the global identifier `birthDate`. That means, in turn, that information from all these language versions of DB-

[1] See: `http://mappings.dbpedia.org`

[2] See: `http://mappings.dbpedia.org/index.php/Mapping_Statistics`

pedia can be merged. Knowledge bases for smaller languages can therefore be augmented with knowledge from larger sources such as the English edition. Conversely, the larger DBpedia editions can benefit from more specialized knowledge from localized editions (Tacchini et al., 2009).

2.2. The Lexicalization Data Set

DBpedia also provides data sets explicitly created to support natural language processing tasks. The DBpedia Lexicalization Data Set provides access to alternative names for entities and concepts, associated with several scores estimating the association strength between name and URI. Currently, it contains 6.6 million scores for alternative names.

Three DBpedia data sets are used as sources of name variation: Titles, Redirects and Disambiguation Links[3]. *Labels* of the DBpedia resources are created from Wikipedia page titles, which can be seen as community-approved surface forms. *Redirects* to URIs indicate synonyms or alternative surface forms, including common misspellings and acronyms. As redirects may point to other redirects, we compute the transitive closure of a graph built from redirects. Their labels also become surface forms. *Disambiguation Links* provide ambiguous surface forms that are "confusable" with all resources they link to. Their labels become surface forms for all target resources in the disambiguation page. Note that we erase trailing parentheses from the labels when constructing surface forms. For example the label '*Copyright (band)*' produces the surface form '*Copyright*'. This means that labels of resources and of redirects can also introduce ambiguous surface forms, additionally to the labels coming from titles of disambiguation pages. The collection of surface forms created as a result of this step constitutes an initial set of name variations for the target resources.

We augment the name variations extracted from titles, redirects and disambiguations by collecting the *anchor texts* of page links on Wikipedia. Anchor texts are the visible, clickable text of wiki page links that are specified after a pipe symbol in the MediaWiki syntax (e.g. `[[Apple_Inc.|Apple]]`). By collecting all occurrences of page links, we can create statistics of co-occurrence for entities and their name variations. We perform this task by counting how many times a certain surface form sf has been used to link to a page uri. We calculate the conditional probabilities $p(uri|sf)$ and $p(sf|uri)$ using maximum likelihood estimates (MLE). The pointwise mutual information $pmi(sf, uri)$ is also given as a measure of association strength. Finally, as a measure of the prominence of a DBpedia resource within Wikipedia, $p(uri)$ is estimated by the normalized count of incoming page links of a uri in Wikipedia.

This data set can be used to estimate ambiguity of phrases, to help select unambiguous identifiers for ambiguous phrases, or to provide alternative names for entities, just to cite a few examples. The DBpedia Lexicalization Data Set has been used as one of the data sources for developing DBpedia Spotlight, a general-purpose entity disambiguation system (Mendes et al., 2011b).

[3]http://wiki.dbpedia.org/Downloads37

```
1 dbpedia:Alkane  carbon alkanes atoms
2 dbpedia:Astronaut space nasa
3 dbpedia:Apollo_8  first moon week
4 dbpedia:Actinopterygii  fish species genus
5 dbpedia:Anthophyta  forests temperate plants
```

Figure 1: A snippet of the Topic Signatures Data Set.

By analyzing the DBpedia Lexicalization Data Set, one can note that approximately 4.4 million surface forms are unambiguous and 392,000 are ambiguous. The overall average ambiguity per surface form is 1.22 – *i.e.* the average number of possible disambiguations per surface form. Considering only the ambiguous surface forms, the average ambiguity per surface form is 2.52. Each DBpedia resource has an average of 2.32 alternative names. These statistics were obtained from Wikipedia dumps using a script[4] written in Pig Latin (Olston et al., 2008) which allows its execution in a distributed environment using Hadoop[5].

2.3. The Topic Signatures Data Set

The Topic Signatures Data Set enables the description of DBpedia Resources in a more unstructured fashion, as compared to the structured factual data provided by the Mapping-based properties. We extract paragraphs that contain wiki links to the corresponding Wikipedia page of each DBpedia entity or concept. We consider each paragraph as contextual information to model the semantics of that entity under the Distributional Hypothesis (Harris, 1954). The intuition behind this hypothesis is that entities or concepts that occur in similar contexts tend to have similar meanings. We tokenize and aggregate all paragraphs in a Vector Space Model (Salton et al., 1975) of terms weighted by their co-occurrence with the target entity. In our VSM, each entity is represented by a vector, and each term is a dimension of this vector. Term scores are computed using the TF*IDF weight.

We use those weights to select the strongest related terms for each entity and build topic signatures (Lin and Hovy, 2000). Figure 1 shows examples of topic signatures in our data set.

Topic signatures can be useful in tasks such as Query Expansion and Document Summarization (Nastase, 2008). An earlier version of this data set has been successfully employed to classify ambiguously described images as good depictions of DBpedia entities (García-Silva et al., 2011).

2.4. The Thematic Concepts Data Set

Wikipedia relies on a category system to capture the idea of a 'theme', a subject that is discussed in its articles. Many of the categories in Wikipedia are linked to an article that describes the main topic of that category. We rely on this information to mark DBpedia entities and concepts that are 'thematic', that is, they are the center of discussion for a category.

[4]Script available at `https://github.com/dicode-project/pignlproc`

[5]http://hadoop.apache.org

```
SELECT ?resource
WHERE {
 ?resource dcterms:subject
      <http://dbpedia.org/resource/Category:Biology> .
}
```

Figure 2: SPARQL query demonstrating how to retrieve entities and concepts under a certain category.

```
PREFIX dbpedia-owl:<http://dbpedia.org/ontology/>
PREFIX dbpedia:<http://dbpedia.org/resource/>
SELECT ?resource
WHERE {
?resource dbpedia-owl:wikiPageWikiLink dbpedia:Biology .
}
```

Figure 3: SPARQL query demonstrating how to retrieve pages linking to topical concepts.

A simple SPARQL query can retrieve all DBpedia resources within a given Wikipedia category (Figure 2). A variation of this query can use the Thematic Concepts Data Set to retrieve other DBpedia resources related to certain theme (Figure 3). The two queries can be combined with trivial use of SPARQL UNION. This set of resources can be used, for instance, for creating a corpus from Wikipedia to be used as training data for topic classifiers.

2.5. The Grammatical Gender Data Set

DBpedia contains 416,000 instances of the class Person. We have created a DBpedia Extractor that uses a simple heuristic to decide on a grammatical gender for each person extracted. While parsing an article in the English Wikipedia, if there is a mapping from an infobox in this article to the class dbpedia-owl:Person, we record the frequency of gender-specific pronouns in their declined forms (Subject, Object, Possessive Adjective, Possessive Pronoun and Reflexive) – i.e. he, him, his, himself (masculine) and she, her, hers, herself (feminine).

```
dbpedia:Aristotle foaf:gender "male"@en .
dbpedia:Abraham_Lincoln foaf:gender "male"@en .
dbpedia:Ayn_Rand foaf:gender "female"@en .
dbpedia:Andre_Agassi foaf:gender "male"@en .
dbpedia:Anna_Kournikova foaf:gender "female"@en .
dbpedia:Agatha_Christie foaf:gender "female"@en .
```

Figure 4: A snippet of the Grammatical Gender Data Set.

We assert the grammatical gender for the instance being extracted if the number of occurrences of masculine pronouns is superior than the occurrence of feminine pronouns by a margin, and vice-versa. In order to increase the confidence in the extracted grammatical gender, the current version of the data set requires that the difference in frequency is 200%. Furthermore, we experimented with a minimum occurrence of gender-specific pronouns on one page of 5, 4 and 3. The resulting data covers 68%, 75% and 81%, respectively, of the known instances of persons in DBpedia. Our extraction process assigned the grammatical gender "male" to roughly 85% and "female" roughly 15% of the people. Figure 4 shows example data.

2.6. RDF Links to other Data Sets

DBpedia provides 6.2 million RDF links pointing at records in other data sets. For instance, links to Word Net Synsets (Fellbaum, 1998) were generated by relating Wikipedia infobox templates and Word Net synsets and adding a corresponding link to each entity that uses a specific template. DBpedia also includes links to other ontologies and knowledge bases, including Cyc (Lenat, 1995), Umbel.org, Schema.org and Freebase.com.

Other useful linked sources are Project Gutenberg [6], which offers thousands of free e-books and New York Times, which began to publish its inventory of articles collected over the past 150 years. As of January 2010, 10,000 subject headings had been shared. The links from DBpedia to authors and texts in Project Gutenberg could be used for backing author identification methods, for instance. Meanwhile, the links to concepts in the New York Times database, enable its usage as an evaluation corpus (Sandhaus, 2008) for Named Entity Recognition and Disambiguation algorithms, amongst others.

3. Use Cases

In this section, we outline four use cases of the DBpedia knowledge base in tasks related to computational linguistics and natural language processing.

3.1. Reference Knowledge Base for Disambiguation Tasks

The existence of a homogenized schema for describing data in DBpedia, coupled with its origins on the largest source of multilingual encyclopaedic text available, makes this knowledge base particularly interesting as a resource for natural language processing. DBpedia can be used, for instance, as a reference knowledge base for Entity Linking (McNamee et al., 2010), and other Word Sense Disambiguation-related tasks.

For example, the Entity Linking task at TAC-KBP 2011 (Ji et al., 2011) uses a target knowledge base that can be automatically mapped to DBpedia via Wikipedia links. It has been shown that simple entity linking algorithms can leverage this mapping to obtain a μAVG of 0.827 in the TACKBP-2010 and 0.727 in TACKBP-2011 data sets (Mendes et al., 2011a). A number of academic and commercial projects already perform Entity Linking directly to DBpedia (Mendes et al., 2011b; Ltd., 2009; Reuters, 2008), and others can be mapped to DBpedia via Wikipedia (Orchestr8 LLC, 2009; Giuliano et al., 2009; Ferragina and Scaiella, 2010; Ratinov et al., 2011; Han and Sun, 2011).

One advantage of using DBpedia over Wikipedia as target knowledge base for evaluations is the DBpedia Ontology. By providing a hierarchical classification of concepts, DBpedia allows one to select a subset of classes on which to

[6] See: `http://www.gutenberg.org/`

```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbpedia-owl:<http://dbpedia.org/ontology/>
SELECT DISTINCT ?person
WHERE {
  ?person rdf:type dbpedia-owl:Person.
}
```

Figure 5: SPARQL query demonstrating how to select all instances of type Person.

```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbpedia-owl:<http://dbpedia.org/ontology/>
SELECT DISTINCT ?person ?link
WHERE {
  ?person rdf:type dbpedia-owl:Person .
  ?person dbpedia-owl:wikiPageWikiLink ?link .
}
```

Figure 6: SPARQL query demonstrating how to select all pages linking at entities of type Person.

focus a particular disambiguation task. With a simple Web query (Figure 5) one can obtain a list of entities of type Person or Organization (or even more specific types such as Politician or School).
Simple extensions to those queries can also retrieve a list of all Wikipedia pages that link to entities matching those queries. An example of such a query[7] is shown in Figure 6. These pages, along with the in-text links can be used as training data for Named Entity Recognition or Entity Linking algorithms, for example. A similar approach is used by DBpedia Spotlight.

3.2. Question Answering: World Knowledge

Automatic answering of natural language questions gains importance as information needs of non-technical users grow in complexity. Complex questions have been traditionally approached through the usage of databases and query languages. However, such query languages may not be a viable option for non-technical users. Moreover, alongside structured information in databases, the amount of information available in natural language increases at a fast pace. The complexity of retrieving required information and the complexity of interpreting results call for more than classical document retrieval.
DBpedia contains structured information about a variety of fields and domains from Wikipedia. This information can be leveraged in question answering systems, for example, to map natural language to a target query language. The QALD-1 Challenge (qal, 2011) was an evaluation campaign where natural language questions were translated to SPARQL queries, aiming at retrieving factual answers for those questions. As part of this task, it is necessary to constrain on certain ontology properties (e.g. the gender and age of a person) and it can be beneficial to use the DBpedia ontology. For example, Figure 7 shows the SPARQL query for the question *Who is widow to a politician that died in Texas?* (qal, 2011).

[7]Please note that the wikiPageLinks data set is not loaded in the public SPARQL endpoint, but is available for download and local usage.

```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbpedia-owl:<http://dbpedia.org/ontology/>
PREFIX dbpedia:<http://dbpedia.org/resource/>
SELECT DISTINCT ?widow
WHERE {
  ?politician rdf:type dbpedia-owl:Person.
  ?politician dbpedia-owl:occupation dbpedia:Politician.
  ?politician dbpedia-owl:deathPlace dbpedia:Texas.
  ?politician dbpedia-owl:spouse ?widow.
}
```

Figure 7: SPARQL query demonstrating how to select all pages linking at entities of type Person.

3.3. Slot Filling and Relationship Extraction

Since the DBpedia knowledge base contains also structured information extracted from infoboxes, it can be used as reference knowledge base for other tasks such as slot filling and relationship extraction. Through mappings of several infobox fields to one ontology property, a more harmonized view of the data is provided, allowing researchers to exploit Wikipedia to a larger extent, e.g. attempting multilingual relationship extraction.

3.4. Information Retrieval: Query Expansion

Understanding keyword queries is a difficult task, especially due to the fact that such queries usually contain very few keywords that could be used for disambiguating ambiguous words. While users are typing keywords, current search engines offer a drop-down box with suggestions of common keyword combinations that relate to what the user is typing.
For an ontology-based system that interfaces with the users through keyword searches, such 'auto-suggest' functionality can be achieved through the use of the data in the DBpedia Lexicalization Data Set. Figure 8 shows how to retrieve all resources that are candidate disambiguations for a surface form, along with a score of association strength. The available scores are described in Section 2.2..

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT ?resource ?score WHERE {
GRAPH ?g {
  ?resource skos:altLabel ?label.
}
?g  <http://dbpedia.org/spotlight/score>  ?score.
FILTER (REGEX(?label, "apple", "i"))
}
```

Figure 8: SPARQL query for retrieving candidate disambiguations for the string 'apple'.

4. Conclusion

DBpedia is a multilingual multidomain knowledge base that can be directly used in many tasks in natural language

processing. All DBpedia data sets are freely available under the terms of the Creative Commons Attribution-ShareAlike 3.0 License and the GNU Free Documentation License and can be downloaded from the project website[8]. Furthermore, through the use of W3C-recommended Web technologies, a subset of the DBpedia knowledge base is also available for online usage through Web queries[9].

5. Acknowledgements

We wish to thank Robert Isele and the developers of the DBpedia Extraction Framework, Paul Kreis and the international team of DBpedia Mapping Editors, as well as Dimitris Kontokostas and the DBpedia Internationalization team for their invaluable work on the DBpedia project.

This work was partially funded by the European Commission through FP7 grants LOD2 - Creating Knowledge out of Interlinked Data (Grant No. 257943) and DICODE - Mastering Data-Intensive Collaboration and Decision Making (Grant No. 257184).

6. References

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

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1625–1628, New York, NY, USA. ACM.

Andrés García-Silva, Max Jakob, Pablo N. Mendes, and Christian Bizer. 2011. Multipedia: enriching DBpedia with multimedia information. In *Proceedings of the sixth international conference on Knowledge capture*, K-CAP '11, pages 137–144, New York, NY, USA. ACM.

Claudio Giuliano, Alfio Massimiliano Gliozzo, and Carlo Strapparava. 2009. Kernel methods for minimally supervised wsd. *Comput. Linguist.*, 35:513–528, December.

Xianpei Han and Le Sun. 2011. A generative entity-mention model for linking entities with knowledge base. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 945–954, Portland, Oregon, USA, June. Association for Computational Linguistics.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Heng Ji, Ralph Grishman, and Hoa Dang. 2011. Overview of the TAC2011 Knowledge Base Population Track. In *Proceedings of the Text Analysis Conference (TAC 2011)*.

Douglas Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, November.

Chin-Yew Lin and Eduard H. Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *COLING*, pages 495–501.

Zemanta Ltd. 2009. Zemanta api overview. http://www.zemanta.com/api/.

Paul McNamee, Hoa Trang Dang, Heather Simpson, Patrick Schone, and Stephanie M. Strassel. 2010. An evaluation of technologies for knowledge base population. In *LREC*. European Language Resources Association.

Pablo N. Mendes, Joachim Daiber, Max Jakob, and Christian Bizer. 2011a. Evaluating dbpedia spotlight for the tac-kbp entity linking task. In *Proceedings of the TAC-KBP 2011 Workshop*.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011b. DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*.

Vivi Nastase. 2008. Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 763–772, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins. 2008. Pig latin: a not-so-foreign language for data processing. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, SIGMOD '08, pages 1099–1110, New York, NY, USA. ACM.

Orchestr8 LLC. 2009. AlchemyAPI. `http://www.alchemyapi.com/`, retrieved on 11.12.2010.

2011. *Proceedings of 1st Workshop on Question Answering over Linked Data (QALD-1), collocated with the 8th Extended Semantic Web Conference (ESWC 2011), Heraklion, Greece*, 6.

Lev-Arie Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *ACL*, pages 1375–1384.

Thomson Reuters. 2008. OpenCalais: Connect. Everything. `http://www.opencalais.com/about`, retrieved on 11.12.2010.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, November.

Evan Sandhaus. 2008. The New York Times Annotated Corpus.

Eugenio Tacchini, Andreas Schultz, and Christian Bizer. 2009. Experiments with wikipedia cross-language data fusion. volume 449 of *CEUR Workshop Proceedings ISSN 1613-0073*, June.

[8] `http://dbpedia.org/downloads`

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