Linking WordNet to DBpedia

Aynaz Taheri
Computer Engineering Dept., Shahid Beheshti University, Tehran, Iran
ay.taheri@mail.sbu.ac.ir

Mehrnoush Shamsfard
Computer Engineering Dept., Shahid Beheshti University, Tehran, Iran
m-shams@sbu.ac.ir

Abstract

In this paper we present the process of matching two important datasets in Linking Open Data (LOD): DBpedia and WordNet 3.0. DBpedia plays the main role in the LOD cloud. It is an influential knowledge base and consists of over one billion pieces of information about millions of things. On the other hand Princeton WordNet is the most important lexical ontology. WordNet 3.0 in RDF is also one of the resources in Linking Open Data. Certainly, linking these two common datasets has impressive effects in various aspects of consuming them. In this paper the methodology of matching, statistical information and some beneficial use cases of the matching is explained.

1 Introduction

Nowadays increasing the amount of linked data in Linking Open Data project is not the only challenge of publishing linked data; rather, mapping and linking the linked data resources are also equally important and can improve the effective consuming of linked data resources. Without these links, we confront with isolated islands of datasets, which could not exploit knowledge of each other. The fourth rule of publishing linked data in (Bizer, Heath, et al., 2009) explains the necessity of linking URIs to each other. Therefore, extension of datasets without interlinking them is against the Linked Data principles. The importance of this issue increased our motivation of doing mapping between two core datasets of Linking Open Data.

DBpedia is a significant knowledge base. DBpedia knowledge extraction framework extracted its knowledge from Wikipedia and converted itself as a crystallization point for the web of data (Bizer, Lehmann, et al., 2009). DBpedia currently has knowledge for more than 3.6 million things about persons, places, music, films, video games and etc ("DBpedia," n.d.). It also contains information for these things in different languages. Most of the dataset publishers try to link their datasets to DBpedia. In (Cyganiak, 2010) you can see the mass of links to DBpedia. Some of links from DBpedia to these datasets are available in ("Interlinking DBpedia," 2011) and one of these datasets is WordNet (W3C).

WordNet (Fellbaum,1998) is an electronic lexical database that is designed in Princeton University for English language. WordNet uses synonymous sets, called synset. The latest version of WordNet contains 155,287 words organized in 117,659 synsets ("WordNet 3.0 database," n.d.). WordNet includes nouns, adjectives, verbs and adverbs. Synsets in WordNet are connected to each other with semantic relation such as: synonymy, antonymy, hyponymy, hypernymy, meronymy, troponymy and etc.

Lexical ontologies like WordNet are important resources in natural language processing (NLP). They are used in various tasks and applications, especially where semantic processing is evolved such as question answering, machine translation, text understanding, information retrieval and extraction, knowledge acquisition and semantic search engines (Shamsfard, 2008). Integration of Princeton WordNet and DBpedia could improve the semantic processing. Princeton Wordnet has been mapped to most of the WordNets developed for other languages in the world. So, WordNets of these languages could be linked to DBpedia via Princeton WordNet and the result of WordNet to DBpedia matching will affect NLP in different languages.

At the present time, WordNet is available in

the Linking Open Data cloud. There are two datasets in the LOD cloud which represents WordNet in the form of linked data. One of them is WordNet (W3C[1]) (Assem et al., 2006) that is the OWL/RDF representation of Princeton WordNet 2.0 and the other one is WordNet (VUA[2]) ("Wordnet 3.0 in RDF" 2010) that is the RDF version of WordNet 3.0. WordNet (VUA) is mapped to WordNet (W3C) and DBpedia is also linked to WordNet (W3C).

Each synset in WordNet (VUA) has an URI. Synsets are derefrencable by their URIs and via HTTP protocol. Instances of synsets have also URIs. There are specified patterns for URIs of synsets and instances. The word "instance" is depicted in URIs of instances.

Currently, DBpedia has 467101 links to WordNet (W3C). But there are shortcomings in these links. We are going to cover these defects in our matching:

1. There are only hypernymy relations from instances of DBpedia to noun synsets of WordNet in current links and there is no relation from WordNet to DBpedia. It is considerable that these relations only represent a kind of instantiation. An example of these relations is following:
<http://dbpedia.org/resource/White_House>
< http://dbpedia.org/property/wordnet_type>
<http://www.w3.org/2006/03/wn/wn20/instances/synset-building -noun-1>
In the above example, it is demonstrated that 'White_house' is an instance of 'building' synset in WordNet and nothing more. Whereas, WordNet has information about 'white_house'. There is a synset in WordNet 2.0 with this URI: "http://www.w3c.org/2006/03/wn/wn20/instances/synset-White_house-noun-2". Matching the WordNet synset and the correspondent one in DBpedia is desirable for us. There are many synsets in WordNet which have equivalents in DBpedia. Discovering this type of relations is one of our motivations for doing this project.
2. There is another kind of relation that detecting it between WordNet and DBpedia is beneficial. We find instantiation or hypernymy relations between noun synsets of WordNet and DBpedia classes. It is important to know a synset of WordNet belongs to which concept from the viewpoint of DBpedia.
3. There are many properties in DBpedia. There is no link from these properties to equivalents in WordNet.
4. Only noun synsets of WordNet are considered in current links of DBpedia to WordNet. We are going to find equivalents of verb and adjective synsets in WordNet to properties in DBpedia too.

The rest of the paper is organized as follows: Section 2 presents our methodology of matching WordNet 3.0 in RDF to DBpedia. Section 3 explains statistical information about the result of mapping. Section 4 describe some use cases and advantages of the mapping WordNet3.0 to DBpedia. Section 5 discusses evaluation of the results and section 6 provides some conclusion about this paper.

2 Methodology of Matching

We are going to find coreferent URIs in WordNet (VUA) and DBpedia. This process is also known as entity matching, object resolution, object consolidation, entity identification, identity recognition, identity disambiguation or instance matching. In recent years many efforts have been done for making tools, softwares and frameworks for detecting coreferent URIs. One of these important products is Silk framework (Volz et al., 2009). Silk is a link discovery framework for the web of data that uses a declarative language (Silk-LSL) for specifying which types of links should be found between which types of entities. DBpedia has counseled utilizing this tool for generating links from other datasets to DBpedia ("Interlinking DBpedia," 2011). In our work, we do not use Silk because it needs all kinds of relations that might be discovered between entities to be described by user beforehand. We instead, apply our approach for generating links between two datasets.

Our methodology consists of two main phases: preliminary, supplementary.

1. Preliminary Phase:

In this phase, we use a terminological method for comparing synsets in WordNet and entities in DBpedia. The terminological method is applied in three steps:

- *Matching instances of WordNet to instances and classes of DBpedia:*
  At the first step, equivalent instances in WordNet and DBpedia are found. Instances' synsets in WordNet (UVA) are available apart from other noun synsets. Thus, in the first step we discover all equivalent URIs from these two sets. After finding these equivalences, the next step is to detect the correspondences between Wordnet instances and DBpedia

[1] World Wide Web Consortium
[2] Vrije University Amsterdam

classes.

There are instantiation relations between instances and their classes in DBpedia. In fact, the types of instances are described with these relations. This relation in DBpedia is represented with: "http://www.w3.org/1999/02/22-rdf-syntax-ns#type". Since, we found equivalences relations between instances in WordNet and DBpedia, and on the other hand there are instantiation or hypernymy relations between instances and their classes in DBpedia, so it is possible to represent the type of WordNet instances in DBpedia. Figure 1 indicates this process.

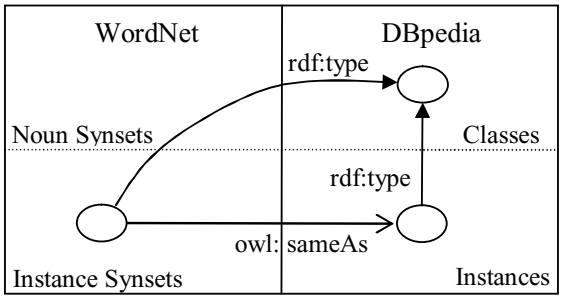

Figure 1. Process of matching at the first step in the first phase

- *Matching noun synsets of WordNet to classes in DBpedia:*

  Some noun synsets of WordNet have equivalents in classes of DBpedia. For example, both of the datasets have knowledge about "Language". So, the "synset-Language-noun-1" synset in WordNet is the same as "Language" class in DBpedia.

  <http://purl.org/vocabularies/princeton/wn30/synset-language-noun-1>
  <http://www.w3.org/2002/07/owl#owl:equivalentClass>
  <http://dbpedia.org/ontology/Language>

- *Matching noun, verb and adjective synsets of WordNet to properties in DBpedia:* The aim of this phase is to recognize properties of DBpedia and their equivalents in WordNet. Properties play the predicate role in a triple. So, detecting equivalent synsets with properties is advantageous. There are three kinds of properties in DBpedia: owl:ObjectProperty, owl:DatatypeProperty and property. owl:ObjectProperty and owl:DatatypeProperty are properties in ontology of DBpedia but the third kind of properties are independent from DBpedia ontology and there are no structural and hierarchical relations between them ("The DBpedia Data Set," 2011). These properties are created directly from Wikipedia infobox properties with no regard to DBpedia ontology. There are some properties in these three kinds that seem to be equivalent. The next example represents that there are two properties in DBpedia that specify "Language" property:

  < http://dbpedia.org/ontology/language> : an object property

  <http://dbpedia.org/property/language> : a property

In all of the three steps, similarity computing method is a token-based distance computing. In this method a string is considered as a bag of words (Euzenat,2007). In our matching method the label of entities in DBpedia, the label of the WordNet synsets in their URIs, the label of senses in a synset and the description (gloss) of a synset in WordNet are transformed to bags of words. But before this transformation we normalize strings and remove stop words. After producing the bags of words, a measure for estimation of similarity between WordNet synset and DBpedia entity is applied (1).

X: a bag of words including words in the label of synset

Y: a bag of words including words in the label of DBpedia entity or its comment

S: a bag of words including words in the label of senses of synset

G: a bag of words including words in the glossary of synset

$$\delta(X,Y) = \frac{|X \cap Y| + |Y \cap (XUSUG)|}{|X| + |Y|} \quad (1)$$

2. Supplementary Phase:

Similar entities regarding lexical features were found in the previous phase. So, we are sure that the matching results of the first phase are lexically correct. But these results are not accurate necessarily and maybe there are correspondences that don't have entities with the same identity and there are only lexical similarities between them. The purpose of this phase is to refine the result of matching in the previous phase. In other words, a method for URI disambiguating is described in this phase. We used hierarchal structure of WordNet and DBpedia for disambiguation. 'wnschema:instanceOf' and 'wn20schema:hyponymOf' relations are used for gaining an understanding of taxonomic structure in WordNet. 'wnschema:instanceOf' relation denotes a relation between an instance synset and a noun synset. Two relations from DBpedia are also utilized for determining the taxonomic structure. These two relations are 'rdf:type' and

'rdfs:subClassOf'. The first one denotes a relation between an instance and the class that belong to and the second one expresses a relation between two classes. These relations are used as sources for disambiguation.

Some structure based techniques are presented in (Euzenat and Shvaiko, 2007). One of them is Wu-Palmer similarity measure (Wu and Palmer, 1994). This similarity measure is used in the second phase of our methodology. Consider the following URIs:

URI1: http://purl.org/vocabularies/princeton/wn30/s ynset-Pluto-noun-1

URI2: http://purl.org/vocabularies/princeton/wn30/synset-Pluto-noun-2

URI3: http://purl.org/vocabularies/princeton/wn30/synset-Pluto-noun-3

URI4: http://dbpedia.org/resource/Pluto

URI1, URI2 and URI3 from WordNet are lexically similar to each other and the first phase finds them equal to URI4 from DBpedia. While, URI1 has 'wnschema:instanceOf' relation with 'fictional_character' and is a cartoon character. URI2 has 'wnschema:instanceOf' relation with 'Greek_deity' and is the god of the underworld in ancient mythology. URI3 has 'wnschema:instanceOf' relation with 'outer_planet' and is a small planet. URI4 has 'rdf:type' relation with 'Planet' class of DBpedia ontology. The matching of URI1 and URI2 with URI4 is obviously wrong.

We use Wu-Palmer measure and assess the similarity of taxonomic structure of URIs in WordNet. In the former example, the similarity of (fictional_character, Planet) is 0.48 and the similarity of (Greek_deity, Planet) is 0.46. These similarities are less than our threshold. Therefore, the matching of URI1 and URI2 with URI4 are excluded from the results.

All of the correspondences with structural similarity less than the threshold are removed from the result of matching.

The result of matching is available at:
http://step1.nlplab.sbu.ac.ir/wordnetdbpedia/matching.aspx

3 Statistical Information of the Matching Outcome

In table 1 the result of WordNet to DBpedia is represented. In the first column, the kinds of synsets are denoted. In the second column the kinds of relations that are discovered and in the third column elements of DBpedia in matching are represented.

4 Use Cases

Influences of the matching consequences are clearly perceptible in the Natural Language Processing and Semantic Processing domains. We categorize use cases of WordNet 3.0 to DBpedia matching in three groups:

- *Enriching WordNet:* Princeton WordNet can be enriched with more relations. Properties in DBpedia can be used for finding more relations in Wordnet.
- *Developing Formal Ontologies*: Princeton WordNet is a lexical ontology and is far from a formal ontology. The types of relations in Word Net are restricted to synonymy, antonymy, hyponymy, hypernymy, meronymy, troponymy. For moving toward a formal ontology, it is necessary to augment the relations between synsets. With utilizing interlinking of WordNet and DBpedia, it is possible to discovering relations between synsets of WordNet via their correspondent entities in DBpedia. Due to the fact that DBpedia is an important knowledge base and have information for entities in the form of properties. These properties can be exploited for making WordNet a formal ontology.
- *Semantic Search:* In semantic search, disambiguating words is a main challenge. Taking advantage of WordNet to DBpedia matching, could help disambiguating through the large amounts of information about entities and properties in DBpedia.
- *Finding more instances for WordNet synsets:* DBpedia is greater than WordNet in the number of instances. We discovered equivalent relations

Subject(WordNet)	Predicate	Object(DBpedia)	Number of Matching
Instances	owl:SameAs	Instances	27923
Instances	rdf:type	Classes	18555
Noun, Verb, Adjective	owl:equivalentProperty	Object Property	583
Noun, Verb, Adjective	owl:equivalentProperty	DataType Property	438
Noun, Verb, Adjective	owl:equivalentProperty	Property	10379
Noun	owl:equivalentClass	Classes	344

Table 1. Result of Matching

between some synsets of WordNet and classes of DBpedia. So we can apply the instances of equivalent class for the instantiation of the synset.

5 Evaluation

We evaluated a subset of matching result manually. This subset contained 500 members. The value of obtained precision is 0.92. Computation of recall is not possible for this project as there is no golden standard and manual extraction of all possible links between these two sets is almost impossible.

6 Conclusions and Future Work

In this paper we discussed about matching two important datasets in LOD cloud: DBpedia and WordNet. Shortcomings about the current links form DBpedia to WordNet presented and the necessity of generating more different kinds of links between them is explained. Interlinking these two datasets can improve applications on natural language processing and semantic processing; furthermore WordNet is also impressionable from the matching result and can move toward an enriched lexical ontology or even a formal ontology.

Future work will focus on mapping WordNets of other languages especially those with less resources to linked data. Linking WordNets to DBpedia is possible via the outcome of Princeton WordNet to DBpedia matching. For example FarsNet, the Persian wordnet (Shamsfard, et al., 2010) is a good candidate for this mapping. FarsNet to Princeton WordNet mapping is available, so matching FarsNet to DBpedia is possible. After linking FarsNet to DBpedia, we are going to extend relations in FarsNet with utilizing DBpedia. Accordingly, FarsNet will be connected to LOD cloud and causes improvements in Persian language semantic processing.

References

Mark V. Assem, Aldo Gangemi and Guus Schreiber. 2006. Conversion of WordNet to a standard RDF/OWL representation. In proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy.

Mark V. Assem, Aldo Gangemi and Guus Schreiber. RDF/OWL Representation of WordNet. (2006). Retrieved April, 2011, from http://www.w3.org/TR/ wordnet-rdf/

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Soren Auer, Christian Becker, Richard Cyganiak and Sebastian Hellmann. 2009. DBpedia- A crystallization point for the Web of Data. *J. Web Sem.* 7(3): 154-165.

Christian Bizer, Tom Heath and Tim Berners-Lee. 2009. Linked Data-The Story So Far, *Int. J. Semantic Web Inf. Syst.* 5(3) :1-22.

Richard Cyganiak and Anja Jentzsch. The Linking Open Data cloud diagram. (2010). Retrieved April, 2011, from http://richard.cyganiak.de/2007/10/lod/

DBpedia. (2011). Retrieved April, 2011, from http://dbpedia.org/About .

Jerome Euzenat and Pavel Shvaiko. 2007. *Ontology matching*. Springer-Verlag, Berlin Heidelberg.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.

Interlinking DBpedia with other Data Sets. (2011). Retrieved April, 2011, from http://wiki.dbpedia.org/ Interlinking.

Mehrnoush Shamsfard, Akbar Hesabi, Hakimeh Fadaie, Niloofar Mansoory, Ali Famian, Somayeh Bagherbeigi, Elham Fekri, Maliheh Monshizadeh, S. Mostafa Assi. 2010. Semi Automatic Development of FarsNet: The Persian WordNet, In Proceedings of the 5th Global WordNet Conference, Mumbai, India.

The DBpedia Data Set. (2011). Retrieved April, 2011, from http://wiki.dbpedia.org/Datasets.

Julius Volz, Christian Bizer, Martin Gaedke, Georgi Kobilarov. 2009. Silk – A Link Discovery Framework for the Web of Data. In Proceedings of the 2nd Workshop about Linked Data on the Web. Madrid, Spain.

WordNet 3.0 database statistics. Retrieved April, 2011, from http://wordnet.princeton.edu/ wordnet/man/wnstats.7WN.html.

Wordnet 3.0 in RDF. (2010). Retrieved April, 2011, from http://semanticweb.cs.vu.nl/lod/wn30/

Zhibiao Wu and Martha Stone Palmer. 1994. Verb Semantics and Lexical Selection. In proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL), 133–138, Las Cruces.