ONTOLOGIES DERIVED FROM WIKIPEDIA
A Framework for Comparison

Alejandro Metke-Jimenez, Kerry Raymond and Ian MacColl
Faculty of Science and Technology, Queensland University of Technology, Brisbane, Australia

Keywords: Ontology, Wikipedia.

Abstract: Since its debut in 2001 Wikipedia has attracted the attention of many researchers in different fields. In recent years researchers in the area of ontology learning have realised the huge potential of Wikipedia as a source of semi-structured knowledge and several systems have used it as their main source of knowledge. However, the techniques used to extract semantic information vary greatly, as do the resulting ontologies. This paper introduces a framework to compare ontology learning systems that use Wikipedia as their main source of knowledge. Six prominent systems are compared and contrasted using the framework.

1 INTRODUCTION

Ontologies play a central role not only in the Semantic Web vision but also in other fields such as natural language processing and document classification. However, there have been several criticisms of the traditional, engineering-oriented approach to ontology building and, therefore, considerable research has been done in the area of automatic ontology learning (Hepp, 2007).

Many researchers have started using Wikipedia as the main source for ontology learning systems. Wikipedia is an interesting resource because it is built by the community so its overall structure has not been imposed but, rather, reflects the consensus reached by its users. Several approaches have been used to extract semantic information from Wikipedia and the type and quality of the ontologies that have been generated varies greatly. This paper addresses the research question of how to compare these approaches.

Section 2 introduces the framework and explains its dimensions in detail. Section 3 provides a description of the six systems that were selected for this analysis that use Wikipedia as their main source of information and shows how these can be classified using the framework. Section 4 discusses future work. Finally, Section 5 summarises the paper's research contributions.

2 A FRAMEWORK FOR COMPARISON

Other frameworks to compare ontology learning systems have been developed in the past (Shamsfard and Abdollahzadeh Barforoush, 2003; Zhou, 2007). However, their dimensions fail to consider key aspects that need to be taken into consideration when dealing with a resource such as Wikipedia. The purpose of the proposed framework is to enable an easy way to compare current approaches that rely specifically on Wikipedia. We propose a framework with the following eight dimensions:

1. Type of ontology being generated
2. Wikipedia features used
3. Derived ontology elements
4. Additional sources used
5. Extraction mechanism
6. Natural language independence
7. Degree of automation
8. Evaluation method

Each dimension is explained in detail in the rest of this section.

2.1 Type of Ontology being Generated

The term *ontology* comes from the field of philosophy but it has been borrowed by the computer science community. Depending on the field, the exact

Metke-Jimenez A., Raymond K. and MacColl I..
ONTOLOGIES DERIVED FROM WIKIPEDIA - A Framework for Comparison.
DOI: 10.5220/0003094003820387
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2010), pages 382-387
ISBN: 978-989-8425-29-4

meaning of ontology can be slightly different. Many researchers refer to the term vaguely, usually citing Gruber's definition that states that an ontology is a "*specification of a set of shared concepts and their relationships in some domain*" (Gruber, 1993). We believe that in the context of the proposed framework, it is necessary to refer to a definition that is much more specific. We propose classifying ontologies based on their goal since we believe this is what defines their characteristics. The following is the classification we use in the framework.

An Association Ontology is the most basic type of ontology and its goal is to provide measures of semantic relatedness between its resources (concepts, individuals, or both). Its relationships have no type but typically include a numeric value to indicate a degree of relatedness between the connected resources.

Lexical Ontologies focus on including several lexical representations of concepts and individuals and also *is-a* relationships between them. These ontologies are used to assist in natural language processing tasks where the word disambiguation problem is a major challenge.

Classification Ontologies are used to classify resources. These ontologies include *is-a* relationships and a large set of concepts and individuals. When a concept in an ontology is used to classify some resource, the user is indicating that the resource is about that concept.

Representational Ontologies are used to provide an abstract representation of some reality. These are usually fine-grained and specific to some domain. Relationships are usually more complex than the ones found in classification ontologies, since modelling a domain often requires expressing additional conditions and restrictions that go beyond the simple taxonomic ones. However, these ontologies do not require complex reasoning since it is not the intention of the modellers to be able to derive new knowledge from the ontology.

Knowledge Repositories are used to answer semantically rich queries. These ontologies enable automated reasoning, comparable to a human expert, and therefore require different types of relationships, axioms, and complex reasoning. In practice, complex reasoning is very expensive computationally so these ontologies are restricted either in size or in complexity.

2.2 Wikipedia Features Used

This dimension is concerned with the Wikipedia features being used as sources of semantic information. A good review of Wikipedia's most important features can be found in (Medelyan et al., 2009).

2.3 Derived Ontology Elements

Ontologies are composed of different elements and also have different levels of expressiveness. Every approach deals at least with the extraction of concepts and relationships between them. For example, a simple way of deriving the concepts of an ontology is by creating one for each article in Wikipedia.

2.4 Additional Sources Used

Even though Wikipedia is a good source of semantic information, many research projects have chosen to use additional sources to improve the quality of the generated ontology.

2.5 Extraction Mechanism

This dimension is concerned with the actual mechanism used to extract the information from Wikipedia (and possibly other sources) and turn it into an ontology. Several different techniques have been used for this purpose and can be classified in the following major categories:

Linguistic Analysis relies on natural language processing and typically do not make use of Wikipedia's special features (it is treated just like regular text).

Rule-based Methods define rules based on common patterns observed in Wikipedia.

Statistical Methods are based on statistical information (such as word co-ocurrence, for example).

Machine Learning Approaches use either supervised or unsupervised machine learning techniques to extract or derive new semantic information.

Connectivity-based Algorithms make use of the internal links in Wikipedia in order to treat it like a graph in which the articles represent the nodes and the links represent the edges.

Transformation-based methods use structured or semi-structured information and define rules to transform it to elements in the target ontology.

2.6 Natural Language Independence

This dimension is closely related to the extraction dimension since it refers to the applicability of the approach to sources in a language different than the one the approach was originally designed to work with.

2.7 Degree of Automation

This dimension is concerned with the degree of manual intervention required for the approach to work and has the following possible values: manual, semi-automatic, or fully automatic.

2.8 Evaluation Method

This dimension classifies the approach used to evaluate the generated ontology. These approaches can be classified in four main categories (Brank et al., 2005):

Comparative methods, where the ontology is compared with a "gold standard".

Proxied methods, where the results of using the ontology in an application domain (such as document classification) are compared.

Data-based methods, where the"fit" of the ontology to a domain is measured from a source of data (such as a collection of documents).

Human-assessed methods, where the quality of the ontology is evaluated by a group of people against some predefined criteria.

3 USING OUR FRAMEWORK

A considerable amount of research has been devoted to the extraction of semantic information from Wikipedia using various approaches (a good review can be found in (Medelyan et al., 2009)). Our framework is useful to classify ontology learning systems that use Wikipedia as their main source of information. Using the framework allows understanding how these systems work and identifying gaps that might be exploited to improve the content of the generated ontologies. This section shows the results of using the framework to classify six systems. Table 1 shows a summary of the results.

3.1 DBpedia

The DBpedia project focuses on extracting simple semantic information from Wikipedia's structure and templates in the form of RDF triples (Auer et al., 2008). The DBpedia dataset contains about 103 million RDF triples. Some of these include very specific information (mainly from the data extracted from the infoboxes) and some include metadata (such as the page links between Wikipedia articles). The dataset is available on the group's web page.

The goal of DBpedia is to create a knowledge repository with general knowledge extracted from Wikipedia. It uses two main sources of information: database dumps and the page templates. The relationships extracted from the database dumps are untyped and only indicate that an article is related somehow to the articles to which it is linked. The templates allow extracting both attributes and several typed relationships, mainly from individuals.

3.2 Wikipedia Thesaurus

The Wikipedia Laboratory group created an association thesaurus from Wikipedia by using several techniques that calculate the semantic relatedness between articles (Ito et al., 2008). The thesaurus is available on the group's web page.

An association thesaurus contains concepts and relationships between them, with a numeric value that indicates how close the concepts are semantically. The researchers use several techniques to achieve the same result. One of these, known as *pfibf* (Path Frequency - Inversed Backward link Frequency), uses Wikipedia's internal links to derive the relatedness measure. It is similar to the traditional *tfidf* (Term Frequency - Inverse Document Frequency) method used in data mining, but specifically designed to deal with Wikipedia's structure. Another method is based on link co-ocurrence analysis. In this approach the relatedness measures are calculated based on the co-occurrence of pairs of links in the articles.

The resulting thesaurus was assessed by humans and also compared with a "gold standard" for word similarity.

3.3 WikiNet

The EML Research group created a large scale, multilingual concept network (Nastase et al., 2010). The resource consists of language-independent concepts, relationships between them, and their corresponding lexical representations in different languages. The dataset is available on the group's web page.

The multi-lingual concept network is similar to WordNet, but derived automatically from Wikipedia. It is not intended to replace WordNet but rather to complement it, since WordNet has limited coverage despite its high quality.

Table 1: Results of applying the framework to the six selected systems.

	System					
Dimension	**DBPedia**	**WikipediaThesaurus**	**WikiNet**	**YAGO**	**WikiOnto**	**WikiTaxonomy**
Type of Ontology	Knowledge repository	Association	Lexical	Knowledge repository	Representational	Classification
Derived Elements	Concepts Individuals Attributes Untyped relations Other relations	Concepts Numeric relations	Concepts Individuals Is-a relations Instance-of relations Other relations	Concepts Attributes Is-a relations Instance-of relations Other relations	Concepts Individuals Is-a relations	Concepts Individuals Is-a relations Instance-of relations
Wikipedia Features	Articles Internal links Templates	Articles Article text Internal links	Articles Categories Internal links Cross-language links Disambiguation pages	Infoboxes Categories Redirects	Articles Internal links Sections	Categories Internal links Article text
Additional Sources	None	None	None	WordNet	WordNet	None
Extraction Mechanism	Transformation-based	Statistical methods Connectivity-based	Linguistic analysis Transformation-based Statistical methods	Rule-based	Transformation-based Machine learning Linguistic analysis	Linguistic analysis Connectivity-based Rule-based
Language Independence	Partial	Yes	Yes	Yes	Partial	Partial
Degree of Automation	Automatic	Automatic	Automatic	Automatic	Assisted	Automatic
Evaluation Method	Not specified	Human assessed Comparative	Human assessed Comparative	Human assessed	Not specified	Comparative

Several Wikipedia features are used to derive the content of the network. Articles and categories are used to derive concepts. In order to derive relations, syntactic analysis is performed on category names to extract relations and these are propagated to their articles. Information found in infoboxes is also used to further type these relationships. Finally, cross-language links are used to create an index that includes the different language representations of each concept.

The resulting concept network was evaluated by comparing it with Cyc and YAGO. The comparison was performed both manually and automatically.

3.4 YAGO

YAGO (Yet Another Great Ontology) is a project that aims to create a huge general-purpose ontology that includes both concepts and named/entities (Suchanek et al., 2008). The ontology can be accessed through several front ends and can also be downloaded from the project's web page.

The authors use different heuristics to extract attributes and relationships from the infoboxes of the articles, deduce types based on Wikipedia's category network, and extract *is-a* relations. Also, WordNet is used as the source for an organised taxonomy in which the concepts extracted from Wikipedia are placed. WordNet *synsets* are used to derive the meaning of Wikipedia concepts, using Wikipedia redirects to find alternative names for entities, and applying heuristics based on the Wikipedia categories to extract additional information.

The resulting ontology was assessed by presenting the facts to human judges for evaluation. For each fact the judges could select if the fact was correct, incorrect, or that they did not know.

3.5 WikiOnto

The goal of the WikiOnto project is not to create an ontology but to provide an environment to assist in the extraction and modelling of ontologies, using Wikipedia as the main source of information (Silva and Jayaratne, 2009). Even though the goal of the project is not to create an ontology, it was included in the evaluation because it can be considered a semi-automatic approach to ontology building.

The environment enables users to choose an article of interest as the starting point and it provides suggestions of other concepts that might be relevant to include in the ontology using clustering techniques. The ontology builder can add new concepts to the ontology and then specify the relationships between them. The authors plan to implement syntactic analysis in order to suggest relations between concepts other than *is-a*.

Table 2: Derived elements and Wikipedia features used in the evaluated systems.

	Derived Element						
			Relations				
Wikipedia Feature	Concepts & Individuals	Attributes	Untyped	Numeric	Is-a	Instance-of	Other
Articles							
Title	D, T, N, Y, O						
Definition Sentence							
Overview Paragraph							
Sections	O				O		
Text				T	X	X	
Templates		D					D
Infoboxes							N, Y
Internal Links	O		D	T	O		
External Links							
Cross-language Links							N
Category Links					N	Y	
Categories							
Title	N, X				N, Y, X	X	
Text							
Parent Categories					N, X	X	
Cross-language Links							N
Redirects							Y
Disambiguation Pages					N		
Edit Histories							
Discussion Pages							

D = DBpedia, T = Wikipedia Thesaurus, N = WikiNet, Y = YAGO, O = WikiOnto, X = WikiTaxonomy

3.6 WikiTaxonomy

WikiTaxonomy is an ontology derived from Wikipedia's category structure. It includes concepts, individuals, and simple taxonomic relations (Ponzetto and Strube, 2008). The ontology is available in RDFS format.

The authors use a combination of connectivity-based methods and linguistic-based methods to derive *is*-a relations between Wikipedia categories. These relations are then propagated using inference-based methods. Finally, concepts and individuals are identified by using several heuristics.

4 FUTURE WORK

By analysing the cross-product between the framework's *derived elements* and *Wikipedia features* dimensions it is easy to identify which Wikipedia features have been used to derive certain type of semantic information. Table 2 shows the result of applying this cross-product to the six systems that were previously classified with the framework. Future work will involve using this information to identify new sources of semantic information and deriving new mechanisms to exploit it.

5 CONCLUSIONS

This paper introduces a framework to compare approaches to deriving ontologies automatically or semi-automatically from Wikipedia. Six prominent systems are classified using the framework and their most relevant characteristics are summarised. Finally, the cross-product of two of the framework's dimensions is used to show how new sources of semantic information in Wikipedia can be identified.

ACKNOWLEDGEMENTS

This research is supported in part by the CRC Smart Services, established and supported under the Australian Government Cooperative Research Centres Programme, and a Queensland University of Technology scholarship.

REFERENCES

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2008). Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, volume

4825/2008, pages 722–735. Springer Berlin / Heidelberg.

Brank, J., Grobelnik, M., and Mladenić, D. (2005). A survey of ontology evaluation techniques. In *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005)*.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 6(2):199–221.

Hepp, M. (2007). Possible ontologies: How reality constrains the development of relevant ontologies. *IEEE Internet Computing*, 11(1):90–96.

Ito, M., Nakayama, K., Hara, T., and Nishio, S. (2008). Association thesaurus construction methods based on link co-occurrence analysis for wikipedia. In *Proceeding of the 17th ACM Conference on Information and Knowledge Management*, Napa Valley, California, USA. ACM.

Medelyan, O., Milne, D., Legg, C., and Witten, I. H. (2009). Mining meaning from wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754.

Nastase, V., Strube, M., Boerschinger, B., Zirn, C., and Elghafari, A. (2010). Wikinet: A very large scale multilingual concept network. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Ponzetto, S. and Strube, M. (2008). WikiTaxonomy: A large scale knowledge resource. In *Proceeding of the 2008 conference on ECAI 2008: 18th European Conference on Artificial Intelligence*, pages 751–752. IOS Press.

Shamsfard, M. and Abdollahzadeh Barforoush, A. (2003). The state of the art in ontology learning: a framework for comparison. *Knowl. Eng. Rev.*, 18(4):293–316.

Silva, L. N. D. and Jayaratne, L. (2009). Wikionto: A system for semi-automatic extraction and modeling of ontologies using wikipedia xml corpus. *International Conference on Semantic Computing*, 0:571–576.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217.

Zhou, L. (2007). Ontology learning: state of the art and open issues. *Information Technology and Management*, 8(3):241–252.