Leopold–Franzens–University
Innsbruck

Institute of Computer Science
STI - Innsbruck

Ontology Learning

Seminar Paper

Applied Ontology Engineering (WS 2010)

Supervisor: Dr. Katharina Siorpaes

Michael Rogger and Stefan Thaler

Innsbruck, December 3, 2010

Ontology Learning

Michael Rogger and Stefan Thaler

michael.rogger@uibk.ac.at and stefan.thaler@uibk.ac.at

1 Introduction

1.1 What is Ontology Learning?

Ontology learning is a subtask of information extraction, which itself is a type of information retrieval. The main goal of ontology learning is extract relevant concepts and relations from a given data source. A data source can be structured, semi structured or not structured. Due to its nature, unstructured data needs a lot of techniques to achieve good results. A data source can be e.g. texts, relational databases, html documents,... In general ontology learning is (semi-)automatic, so user involvement is need in order to achieve good results. This is based on the nature of limited "intelligence" in machines, they lack to understand the intent and the purpose. The process of ontology learning can be divided in four key steps which is shown in Figure1.

1.2 Why is it needed?

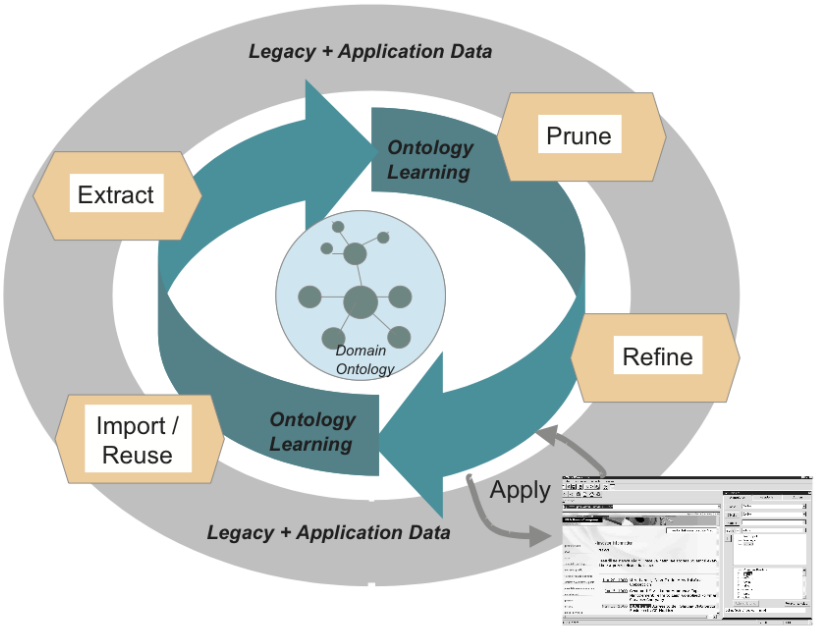
One could argue that ontologies can be built manually by ontology engineers. Human understand the intent and the purpose of a data source (e.g. a text) and can of course derive much better results than machines can. This would suggest automatic methods are not needed. But consider a massive amount of data (e.g. the web), so to a certain degree it is possible to process this manually, but this is rather inefficient with respect to time and may also not be possible with respect to available resources. So this "handwork" can be reduced by using ontology learning techniques, which support the ontology engineer in building these ontologies. So a huge amount of data can be processed in a (semi-)automatic way.

1.3 Why is it difficult?

As mentioned in Section 1.1 the main goal of ontology learning is to extract all relevant parts of an ontology from different data sources. In order to understand the difficulities in that task we need to look at the definition of an ontology and on the given data source.

The definition of an ontology is a "formal and explicit specification of a shared conceptualisation"[2]. So an ontology provides shared vocabulary to model a domain.

Figure 1. Ontology Learning process steps, Figure taken from[1]



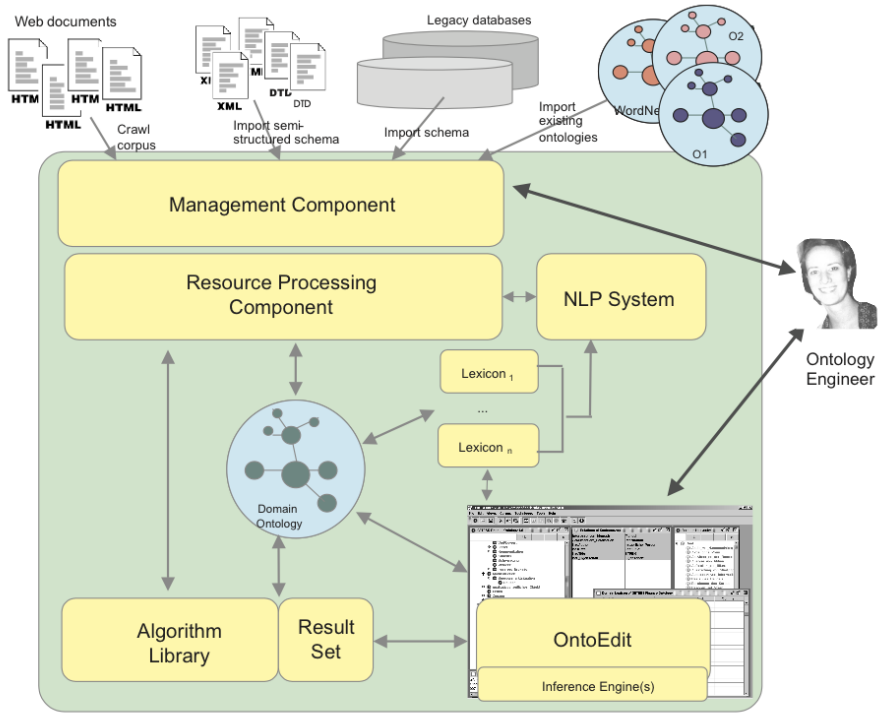
As data source we concentrate on a human written text, because processing unstructured data is the most complicated problem. Brewster et al. [3] argued, that authors writing articles assume a large background knowledge that they share with a community. Authors try to focus in their written texts on specific aspects. Thus most of the knowledge is implicit and allows also to conceptualize this by different people in different manner, even using the same words.

Summarizing it can be said, it is a hard task to close the semantic gap between human language and formalized knowledge because formalized knowledge is declarative and human language is mostly very implicit, vague and defeasible[4].

1.4 Application

The most prominent application of ontology learning might be the semantic web[1]. The semantic web builds on top of ontologies, to be machine understandable. Therefore the success of the semantic web strongly depends on ontologies. Ontology learning provides great facilities to build in a fast and easy way ontologies by an ontology engineer in an (semi-)automatic way. A possible architecture for ontology learning is shown in Figure 2.

Figure 2. Architecture for Ontology Learning for the Semantic Web, Figure taken from[1]



1.5 Exemplary Ontology Learning Methodology

A general approach for how Ontologies can be obtained from non ontological resources has been developed in course of the NeON project[1]. Their methodology is based on three steps. In the first step, the non ontological resource has to be analyzed in order to extract the underlying schema, concepts and data model of the resource. Then, the resource has to be transformed into a conceptual model. This may include automatic tasks as well as manual task. Finally, the conceptual model has to be formalized into an Ontology. Apart from that they provide a collection of guidelines and best practices how the ontology learning process should be conducted [2].

[1] NeON Methodology, http://www.neon-project.org/web-content/media/book-chapters/Chapter-08-1.pdf
[2] Ontology Design Pattern, http://ontologydesignpatterns.org/

2 Ontology Learning from Relational Databases

Relational Databases are widely used in information management, they provide a vast amount of information of different domains in a structured way. There are several approaches to obtain an ontology from a relational database. Most approaches analyze the database scheme and try to deduce additional information about hierachies, properties, ... The work from Man Li et al. [5] is based on a group of learning rules. So these learning rules can detect classes, class hierachies, properties, property characteristics and cardinalities.

3 Ontology Learning from Texts

Ontology learning from text is a very well established research area, as a great share of human knowledge is still available in text form. As mentioned in [6] text sources offer highest availability in contrast to lowest accessibility in terms of mining structured data from it. Consequently, diverse techniques to extract ontologies from text sources have been developed over the past few years. The input for text based ontology learning methods can be distinguished by completely unstructured text or semi structured text such as Web documents. The output of the methods can be grouped together to terms, synonyms, concepts, concept hierarchies, relations and rules as proposed in [7] and shown in figure 3. The field of ontology learning from text deals mainly with learning concepts and their relationships, however to establish this the other elements also have to be extracted from the text source.

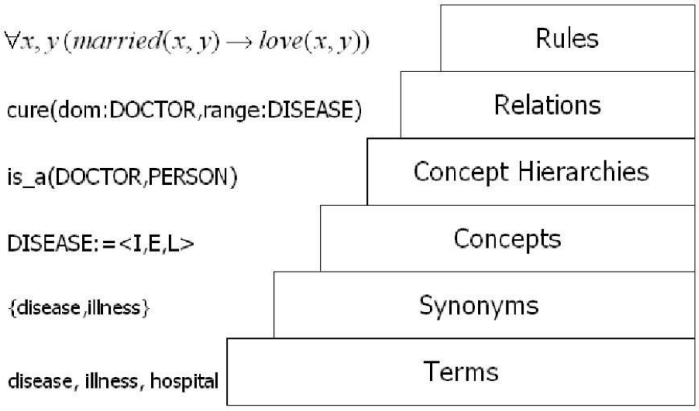


Figure 3. Ontology Learning(from Text) Layer Cake, Figure taken from [7]

Learning terms from text. As stated in [8] a term is "*a word or phrase used to describe a thing or to express a concept, especially in a particular kind of language or branch of study*". In context of ontology learning terms are the groundwork for further analysis. Most techniques belong to the research area of Natural Language Processing, deep language analysis or information retrieval methods for term indexing.

Learning synonyms from text. The Oxford Dictionary defines synonym as "*a word or phrase that means exactly or nearly the same as another word or phrase in the same language*" . To gain synonyms from text mainly existing synonym sources such as WordNet [3] or EuroWordNet[4] and clustering techniques or statistical methods.

Learning concepts from text. Most approaches to aquisit concepts from a text, i.e. abstract ideas that are represented by a charaterizing string tackle the problem from a linguistic point of view. Apart from that formal as well as informal definition are taken into account [7].

Learning concept hierarchies from text. Mining concept taxonomies from text, i.e. a hiarchical is-a relationship between are based on detecting lexico-syntactic patterns, clustering algorithms and term subsumption.

Learning concept relations from text. A greate share of methods that learn non-hierarchical relationships from texts are again based on NLP and statistical techniques. Other, more recent approaches normalize the mentions of concepts and establishing links between them.

Learning rules from text. Finally, acquisiting rules from text mainly focusses on gaining lexical entailments from text. However, this area in ontology learning is not as well established as the the others.

3.1 Text-To-Onto Framework

The Text-To-Onto framework, a tool to extract Ontologies from text has initially been proposed by [9] Its development has been resumed in the NeON Project[5] where it has integrated as a part of the NeON Toolkit. It utilizes a variety of algorithms, measurements and NLP techniques to extract concepts, instances, Subclass-of Relations, Instance-of Relations, General Relations, Equivalences and Subtopic-of relations [9] .

[3] Wordnet, http://wordnet.princeton.edu/
[4] EurWordNet,http://www.illc.uva.nl/EuroWordNet/
[5] NeON project, http://www.neon-project.org/

4 Ontology Learning from Web-Documents

The endless amount of available Web Documents lead to a manifold of methods for ontology learning from them.

On approach, described in [10] defines an algorithm to extract terms, synonyms and semantic relations of Web Documents based on markup exploitation, span counting and frequency order.

ArtEquAKT[6] as described in [11] attemps to implement a system that automatically collects knowledge about artists, extracts ontologies from the documents and then automatically assemble biogrpahies of those artists.

5 Ontology Learning from Folksonomies

Folksonomies, that are *"unsystematic, unsophisticated collections of keywords associated by social bookmarking users to web content"* [12] can be used as a semi structured source for ontology learning. Methods to learn ontologies from Folksomonies differ from others since the regular ontolgy learning processes assume well formed, spelling mistakes free terms. That is not the case with Folksonomies, as they are user created, error prone and may consist of colloquial terms that may not yet be found in dictionaries.

Jie Tang et al. proposes a technique for folksonomy based ontology learning that uses a three phased algorithm [13]. Firstly, they try to extract correspondences between tags and topics. Secondly, they attempt to guess possible relationships between tags and finally this guessed tag relationships are determined and a hierarchy is created.

Another approach suggested by [14] defines a three step process for folksonomy grounded ontology learning. Firstly, they pre-processs existing tags in order to "clean" and sort them. Then they apply clustering algorithms to group these tags and lastly use several measures to identify concepts and relationships between the so generated tag-clusters.

[6] ArtEquAkt, http://www.aktors.org/technologies/artequakt/

References

1. Maedche, A., Staab, S.: Ontology learning for the semantic web. Intelligent Systems, IEEE **16** (2005) 72–79
2. Gruber, T., et al.: A translation approach to portable ontology specifications. Knowledge acquisition **5** (1993) 199–199
3. Brewster, C., Ciravegna, F., Wilks, Y.: Background and foreground knowledge in dynamic ontology construction. In: In Proceedings of the SIGIR Semantic Web Workshop. (2003)
4. Buitelaar, P., Cimiano, P.: Ontology learning and population: bridging the gap between text and knowledge. Ios Pr Inc (2008)
5. Li, M., Du, X., Wang, S.: Learning ontology from relational database. In: Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on. Volume 6., IEEE (2005) 3410–3415
6. Biemann, C.: Ontology learning from text: A survey of methods. LDV Forum **20** (2005) 75–93
7. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer, Berlin (2006)
8. : (Oxford dictionary - term)
9. Cimiano, P., Völker, J.: Text2onto. In: NLDB. (2005) 227–238
10. Brunzel, M., Spiliopoulou, M.: Discovering groups of sibling terms from web documents with xtreem-sg. In Spaccapietra, S., Pan, J., Thiran, P., Halpin, T., Staab, S., Svatek, V., Shvaiko, P., Roddick, J., eds.: Journal on Data Semantics XI. Volume 5383 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2008) 126–155 10.1007/978-3-540-92148-6-5.
11. Alani, H., Kim, S., Millard, D.E., Weal, M.J., Hall, W., Lewis, P.H., Shadbolt, N.: Automatic ontology-based knowledge extraction from web documents. IEEE Intelligent Systems **18** (2003) 14–21
12. Tatu, M., Moldovan, D.: Inducing ontologies from folksonomies using natural language understanding. In Chair), N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D., eds.: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, European Language Resources Association (ELRA) (2010)
13. Tang, J., fung Leung, H., Luo, Q., Chen, D., Gong, J.: Towards ontology learning from folksonomies. In: IJCAI. (2009) 2089–2094
14. Specia, L., Motta, E.: Integrating folksonomies with the semantic web. In: ESWC. (2007) 624–639
15. Chamberlain, J., Poesio, M., Kruschwitz, U.: A demonstration of human computation using the phrase detectives annotation game. In: KDD Workshop on Human Computation. (2009) 23–24
16. Brewster, C., Ciravegna, F., Wilks, Y.: User-centred ontology learning for knowledge management. In: NLDB. (2002) 203–207