Ontologies on Demand?

A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text

Philipp Cimiano, Johanna Völker und Rudi Studer, Karlsruhe

Ontologies are nowadays used for many applications requiring data, services and resources in general to be interoperable and machine understandable. Such applications are for example web service discovery and composition, information integration across databases, intelligent search, etc. The general idea is that data and services are semantically described with respect to ontologies, which are formal specifications of a domain of interest, and can thus be shared and reused in a way such that the shared meaning specified by the ontology remains formally the same across different parties and applications. As the cost of creating ontologies is relatively high, different proposals have emerged for learning ontologies from structured and unstructured resources. In this article we examine the maturity of techniques for ontology learning from textual resources, addressing the question whether the state-of-the-art is mature enough to produce ontologies 'on demand'.

Lernen von Ontologien aus Texten: Stand der Technik, Anwendungen, Herausforderungen und Trends

Ontologien spielen heutzutage eine wichtige Rolle in Anwendungen, die auf die Interoperabilität und Maschinenverständlichkeit von Daten, Diensten und Ressourcen im allgemeinen bauen. Solche Anwendungen findet man z.B. in den Bereichen der Suche und Komposition von Diensten, Integration von Informationen aus verschiedenen Datenbanken und intelligente Suche von Ressourcen. Die generelle Idee ist dabei, dass Daten und Dienste, die semantisch in Bezug zu einer Ontologie beschrieben werden, über verschiedenen Anwendungen und Parteien hinweg mit einer wohldefinierten und auf Konsens beruhenden Bedeutung verwendet werden können, die von der Ontologie formal spezifiziert wird. Da die Erstellung von Ontologien typischerweise relativ kostspielig ist, sind in der Vergangenheit Verfahren zum automatischen Lernen von Ontologien aus strukturierten und unstrukturierten Ressourcen vorgeschlagen worden. In diesem Artikel wird der Reifegrad von Techniken zum Lernen von Ontologien aus textuellen Quellen analysiert, besonders im Hinblick auf die Frage, ob mit solchen Verfahren Ontologien ‚auf Knopfdruck' gelernt werden können.

1 Introduction

The aim of the Semantic Web as originally envisioned by Tim Berners-Lee and others [Berners-Lee et al. 2001] is to add a layer of meaning on top of data, services and resources in general to enforce their interoperability and enable machine interpretability. This is especially important for applications and scenarios in which data needs to be shared in a way such that their meaning is preserved. The data, services and resources are then described semantically via metadata describing their meaning or capabilities. These metadata are captured with respect to *ontologies*, which are logical theories and thus have a formal logical interpretation which is independent of specific applications. Ontologies can then be formalized in different *ontology languages* such as the W3C Standard OWL[1].
If data is annotated with metadata specified with respect to an ontology, it can not only be shared between different parties in a meaning-preserving way, but it can be searched and retrieved in a more effective way. If services are annotated semantically, the search for the appropriate service can be made more effective, and different services can be orchestrated and composed on the basis of their semantic descriptions to achieve a more complex goal.

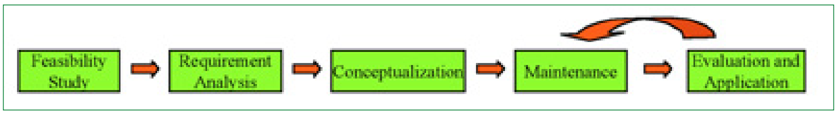

Figure 1: Ontology Engineering Process

A crucial question in the vision of a semantic web is how to support and ease the process of creating and maintaining ontologies. By definition, ontologies represent a shared conceptualization of a domain of interest and should thus be jointly engineered by a number of parties (see the ontology definition in [Gruber 1995]).
As any engineering process, this involves a high effort and some clearly defined methodology consisting at least of a feasibility study, a requirement analysis phase, a conceptualization and formalization phase as well as an application, evaluation and refinement phase (compare Figure 1). It is well known that the process of engineering an ontology is costly. Recently, models for estimating the actual costs have been proposed (compare [Paslaru Bontas et al. 2006]). In order to alleviate the costs involved in the activity of engineering ontologies, several proposals for automatically learning ontologies from data have emerged. In particular, in recent years there has been a great surge of interest in methods for learning ontologies from textual resources, which are still the main means of knowledge externalization used by people.

In this article we will look at the status of maturity of different techniques for learning ontologies from text, examining the maturity of different methods for learning the most important ontological primitives. More importantly, we also discuss potential applications for ontology learning techniques and conclude the article with a reflection on the main challenges and trends within ontology learning. The structure of the article is as follows: in the following section 2 we introduce ontologies as well as ontology engineering, highlighting the role of ontology learning for this purpose. In section 3, we review the state-of-the-art methods applied for learning ontologies from text data and discuss applications for ontology learning in section 4. Before concluding, we summarize the main challenges and trends in the field of ontology learning.

[1] www.w3.org/TR/owl-ref/

2 Ontologies and Ontology Engineering

Ontologies are typically defined as an abstract model of a domain of interest with a formal semantics in the sense that they constitute a logical theory. These models are supposed to represent a shared conceptualization of a domain as they are assumed to reflect the agreement of a certain community or group of people (see [Gruber 1995]). In the simplest case, ontologies consist of a set of concepts or classes[2] which are relevant for the domain of interest as well as a set of relations defined on these concepts. Typically, one distinguishes between taxonomic and non-taxonomic relations. Taxonomic or subclass relations establish a hierarchical order between concepts, which is defined semantically by set inclusion of the members of a class or concept. Non-taxonomic relations are other relations which are defined on concepts but do not serve the purpose of ordering the concepts hierarchically. The concepts and relations can be axiomatically defined by specifying additional properties such as transitivity or cardinality like in database systems.

Ideally, ontologies should be abstract enough in order to be reused across different applications. The crucial question certainly is how such ontologies can be built. The process of creating an ontology is actually an engineering activity the outcome of which is an ontology which represents a shared conceptualization of the people involved in the process. This process is typically referred to as ontology engineering.

Though ontology engineering has been more an art than a science in the past, the Semantic Web community has spent great effort to turn it into the latter by proposing different methodologies and carefully analyzing them. Most of the ontology engineering methodologies distinguish at least the following phases: feasibility study, requirements analysis, conzeptualization and finally deployment, which typically consist in a loop of application, evaluation and maintenance of the ontology (see Figure 1). These phases are sometimes called differently and sometimes partitioned into subphases. Conzeptualization can be for example separated in at least three subtasks: i) development of the domain model, ii) formalization of the model and iii) its implementation in a certain ontology language (see [Pinto & Martins 2004]).

Ontology learning techniques can for example be applied in the conceptualization phase in order to learn a first 'kick-off' model which people can use as a basis for discussion. Once this model is formalized and implemented in a formal ontology language, ontology learning techniques can be applied in the maintenance phase to extend, refine or modify the model.

3 State-of-the-Art in Ontology Learning

In this section we discuss the state-of-the-art in ontology learning by analyzing the most prominent methods applied to learn certain ontology primitives. In particular, we discuss the different methods with respect to the following tasks:

- Extracting the relevant domain terminology and synonyms from a text collection.
- Discovering concepts which can be regarded as abstractions of human thought.
- Deriving a concept hierarchy organizing these concepts.
- Extending an existing concept hierarchy with new concepts.
- Learning non-taxonomic relations between concepts.
- Populating the ontology with instances of relations and concepts.
- Discovering other axiomatic relationships or rules involving concepts and relations.

Table 1 shows different organizations and their ontology learning tools together with the different tasks addressed. We now proceed to discuss the state-of-the-art techniques with respect to the above mentioned tasks, grouping certain tasks wherever appropriate.

3.1 Extracting the Relevant Terminology and Discovery of Synonyms

Classes are the building block of an ontology. In ontology learning, typically the assumption is made that some terms unambiguously refer to a domain-specific concept. Thus, extracting the relevant domain terminology from a text collection is a crucial step within ontology learning. Methods for term extraction can be as simple as counting raw frequency of terms, applying information retrieval methods such as TFIDF (see [Baeza-Yates & Ribeiro-Neto 1999]) or applying sophisticated methods such as the C-value / NC-value method (see [Frantzi & Ananiadou 1999]). In any case, the resulting list of relevant terms will for sure need to be filtered by a domain expert.

In order to detect synonyms, the most common approaches either apply clustering techniques to group similar words together or use some association measure to detect pairs of statistically correlated terms (compare [Manning & Schütze 1999]). The detection of synonyms can help to cluster terms to groups of terms sharing (almost) the same meaning, thus representing ontological classes.

In general, methods for extracting terms and synonyms have gained a reasonable maturity. Synonym extraction methods,

Table 1: Organizations, their systems and the different subtasks addressed taken from [Buitelaar and Cimiano, 2006]

Organization	System	Ontology Learning Subtasks					
		Terms	Synonyms	Concepts	Concept Hierarchy	Relations	Other Axioms
AIFB, Univ. Karlsruhe	TextToOnto/ Text2Onto	X	clusters	X	X	X	X
Amir Kabir Univ., Teheran	HASTI	X			X	X	X
CNTS, Univ. Antwerpen	OntoBasis		clusters	clusters		?	
DFKI	OntoLT / RelExt	X			X	X	
Economic Univ. Prague	TextToOnto Extensions					labels	
ISI , USC	CBC / DIRT		clusters	clusters			
Keio Univ.	DOODLE		similar pairs			X	
Univ. Paris-Sud	ASIUM/ Mo'K		clusters	clusters	X	X	
Univ. Rome	OntoLearn	X	X	X	X	X	
Univ. Salford	ATRACT	X	clusters	clusters			

[2] In the following we will use the terms concept and class synonymously. In the same line, we will regard the notions of relation and property as equivalent.

for example, have been shown to achieve nearly human-like results on the TOEFL synonym task (compare [Turney 2001]).

3.2 Learning Concepts and Concept Hierarchies

The backbone of any ontology is constituted by a set of taxonomic relationships between classes. Each of the classes can be defined intentionally, e.g. by a descriptive label or its relationships to other classes, as well as extensionally by specifying a set of instances belonging to this class. Since the core taxonomy of an ontology, independently of the underlying ontology representation language, is of crucial importance for the use of ontologies as a means of abstraction, most ontology learning approaches so far have focussed on the formation of concepts and concept hierarchies. According to the different ways to define the meaning of classes, various types of systems have recently been developed in the area of ontology learning. The main methods which have been applied to the task of learning subclass relations are unsupervised hierarchical clustering techniques known from machine learning research e.g. [Cimiano et al. 2005, Faure & Nedellec 1999, Caraballo 1999]. These techniques typically learn concepts at the same time as they also group terms to meaning-bearing units which can be regarded as abstractions over words and thus, to some extent, as concepts. Typically, the hierarchies produced by such clustering approaches are very noisy as they highly depend on the frequency and behaviour of the terms in the text collection under consideration. Thus, some researchers have aimed at introducing a supervision into the clustering process by either directly involving the user to validate or reject certain clusters (compare the ASIUM system as described in [Faure & Nedellec 1999]) or including external information to guide the clustering process (see [Cimiano & Staab 2005]).

The other paradigm is due to Marti Hearst and based on the idea that certain patterns reliably indicate a relation of interest between terms. A pattern like "X such as Y" for example indicates that Y is a subclass of X. Though such approaches are more or less reliable, they suffer from a low recall in the sense that such patterns do not occur frequently enough in text data. The proposed solution to this problem is to match these patterns on the Web (see the PANKOW and KnowItAll systems [Cimiano et al. 2004, Etzioni et al. 2004]). Also, other linguistically-inspired heuristics have been applied to increase the coverage of these methods (see [Cederberg & Widdows 2003]). Approaches based on matching such patterns can be implemented relatively easily by using regular expressions and are typically quite efficient as they basically just have to run once through the text collection. The conceptual drawback of such methods is that they essentially discover lexical relations between words but not between concepts, which are supposed to be abstractions and not merely plain words.

Recently, new techniques have been proposed to derive new patterns indicating a relation of interest by bootstrapping procedures which learn new patterns and examples in each iteration (e.g. DIPRE [Brin 1998], Snowball [Agichtein & Gravano 2000], Espresso [Pantel & Pennacchiotti 2006], etc.)

3.3 Extending an Existing Concept Hierarchy with new Concepts

This task consists in extending a concept hierarchy with new concepts by adding a new concept at an appropriate position in the existing taxonomy. Supervised as well as unsupervised methods can be applied for this purpose. In the case of a supervised approach, classifiers need to be trained which predict membership for every concept in the existing concept hierarchy. As such methods need a considerable amount of training data for each concept, such approaches do typically not scale to arbitrary large ontologies. Unsupervised approaches assume a similarity function which computes a measure of fit between the new concept and the concepts existing in the ontology. Such methods rely on an appropriate contextual representation of the different concepts on the basis of which similarity can be computed. In this case, the hierarchical structure of the ontology needs to be considered and somehow integrated into the similarity measure (compare [Cimiano & Völker 2005b] as well as [Pekar & Staab 2003]). Some of these approaches, namely those which build upon a given set of instances of initially unknown classes, inherently tackle both the problem of taxonomy construction and population.

3.4 Learning Non-Taxonomic Relations

Given a taxonomic hierarchy, many existing ontology learning tools try to learn the "flesh" of the ontology, i.e. a set of non-taxonomic relationships which are essential for expressing domain-specific properties of both classes and instances.

In order to learn non-taxonomic relations, one possibility is to learn 'anonymous' associations between terms on the basis of textual material, and then labeling the relations appropriately at a second step. Mädche and Staab for example make use of the well-known association rule learning algorithm to derive such anonymous relations (compare [Mädche & Staab 2000]). Other researchers have mainly exploited verbs appearing in text as indicators of a relation between their arguments (compare [Cimiano et al. 2006], [Ciaramita et al. 2005], [Schutz and Buitelaar 2006]). In general, while the quality of such approaches is in general reasonable, the relations will need to be inspected and validated by an ontology engineer. An important problem for extracting relations is, however, to find the appropriate level of generalization for these relations (compare [Cimiano et al. 2006]).

Depending on concrete applications, other types of relationships such as equivalence, part-of or causality may be of interest. To some extent, one can even consider the identification of meta-properties (properties of classes), as an ontology learning task – although this kind of properties do not form part of standard ontology representation formalisms. A system which automatically learns meta-properties for concepts is the AEON system by [Völker et al. 2005].

The learning of general axioms or rules is currently out of scope for most ontology learning systems, such that we will not discuss this aspect of ontology learning any further.

3.5 Ontology Population

Ontology population essentially consists in adding instances of concepts and relations to the ontology. For the population of ontologies with concept instances, approaches based on matching certain lexico-syntactic patterns - as described above – on the Web using a standard search engine have been shown to perform quite successfully (compare PANKOW [Cimiano et al. 2004] or KnowItAll [Etztioni et al. 2004]). For the task of learning instances of relations, mainly bootstrapping approaches harvesting relation tuples on the Web have been explored (see DIPRE [Brin 1998], Snowball [Agichtein & Gravano 2000], Espresso [Pantel & Pennacchiotti 2006]). In general, it seems that approaches to ontology population have gained a certain maturity and perform reasonably well. We thus conclude that ontology population seems an easier task than the one of learning the actual schema of the ontology.

4 Applications

Currently, there seems to be a trend in the Semantic Web community to focus on what Gruber (compare [Gruber 2004]) called "semi-formal" ontologies. Though there is no clear definition of what a semi-formal ontology is supposed to be, the intuition is that we are talking about an ontology which is to a large extent not axiomatized in the sense of a logical theory. Such ontologies typically consist of a set of concepts and a loosely defined taxonomic organization of these concepts. Such semi-formal ontologies have the potential of providing a benefit for applications which need some abstraction over plain words but do not mainly rely on logical reasoning. Such applications can be mainly found in the fields of information retrieval, text mining and machine learning, where the applied methods are inherently fuzzy and error-prone. In the remainder of this section we examine the application of semi-formal ontologies to the tasks of structuring information as well as information retrieval, but also text mining in general. Further, we discuss a concrete application study to the British Telecom digital library.

4.1 Structuring Information for Advanced Search

As the amount of information available in companies' intranets steadily increases, the need for advanced search functionality grows at the same pace. However, it seems that advanced search functionality presupposes that information is accordingly structured with respect to certain categories. Search tools can then profit of information resources indexed with respect to these categories to provide advanced search functionality such as: search by category, retrieval of documents with similar categories, etc.

Ontology learning can be applied for example to automatically derive categories from the underlying data which can then be used to index information resources and consequently to foster their search and reuse. The indexing of resources with respect to a given taxonomy of categories can either be done manually or by applying text mining techniques. As described below, it has recently even been shown that text mining techniques such as text classification or clustering can also be improved by integrating ontological information.

4.2 Information Retrieval and Text Mining

One of the main problems that people have been struggling with in information retrieval is the so called *vocabulary mismatch problem*, which essentially consists in the fact that, in many cases, the words used in a query do not match the words in a document though both fit each other from a semantic point of view. In text mining, we face a similar situation. Unsupervised techniques used for clustering documents and relying on the bag-of-words model presuppose a considerable overlap in the words contained in documents in order to group them into one cluster. However, in many cases semantic overlap is not reflected in a corresponding overlap of the words in a document. Thus, many proposals in information retrieval and text mining have come up with the idea of integrating hierarchical organizations of words to either expand the query of a user with relevant words or extend bag-of-words approaches by generalizing words along a taxonomic hierarchy. Recent results in the field of classification and clustering of documents have shown that automatically learned concept hierarchies can be successfully applied to partially overcome the vocabulary mismatch problem (compare [Bloehdorn et al. 2005]).

4.3 BT Digital Library Case Study

The 'Digital Library Case Study' at British Telecom is one of three case studies within the EU IST integrated project Semantically Enabled Knowledge Technologies (SEKT[3]) which aims at the development and exploitation of technologies to support the 'Next Generation Knowledge Management' (see [Davies et al. 2006] for details). The library consists of an extensive on-line collection of technical articles, business journals, proceedings and books which are accessible to a variety of different users. The contents are structured by means of 'Information Spaces', i.e. keyword-based queries and associated documents belonging to one or more topics of interest that are defined in a global topic hierarchy.

One of the main objectives of the SEKT case study is to enhance knowledge access to BT's digital library. Up to now, searching and browsing has been limited to simple keyword-based queries and rather broad topics. Four major use cases for ontology learning techniques emerged from the requirement of improved searching and browsing: First, the extraction of ontologies from particular information spaces is supposed to enable a more fine-grained representation of the contents that could be used to improve visualization and browsing of information spaces. Second, topics and relationships learned from the documents may be used for refining the global topic hierarchy. Furthermore, structured knowledge which has been extracted from a particular information space might serve as a basis for semantics-based query expansion. And finally, ontology learning techniques will enable sophisticated knowledge access by means of ontology-based question answering.

Research within the SEKT project has led to the development of a prototype (depicted by Figure 2) which allows for querying the whole variety of information sources provided by the digital library - full text documents, structured metadata and topic hierarchies - by means of a single natural language query. The query is transformed into a structured logical form which is then passed on to a Description Logics reasoner, i.e. KAON2 [Motik & Sattler 2006], that tries to provide an answer by inferencing over an integrated ontology built from the different information sources. Whereas the integration of structured metadata and topic hierarchies into this ontology can be done in a relatively straightforward way, the creation of ontological data from the full text documents requires the application of ontology learning techniques such as those provided by Text2Onto (see [Cimiano & Völker 2005b]). Ontology learning tasks being performed in this scenario include the extraction of concepts, instances as well as taxonomic and non-taxonomic relationships such as part-of and subtopic-of relations.

5 Challenges and Trends

The main trends which can be identified in the field of ontology learning are on the one hand the creation of flexible ontology learning frameworks and the treatment of the uncertainty of the predictions of ontology learning algorithms. Current ontology learning frameworks such as Text2Onto[4] [Cimiano & Völker 2005b] or JATKE[5] have been designed with a central management component allowing various algorithms for ontology learning to

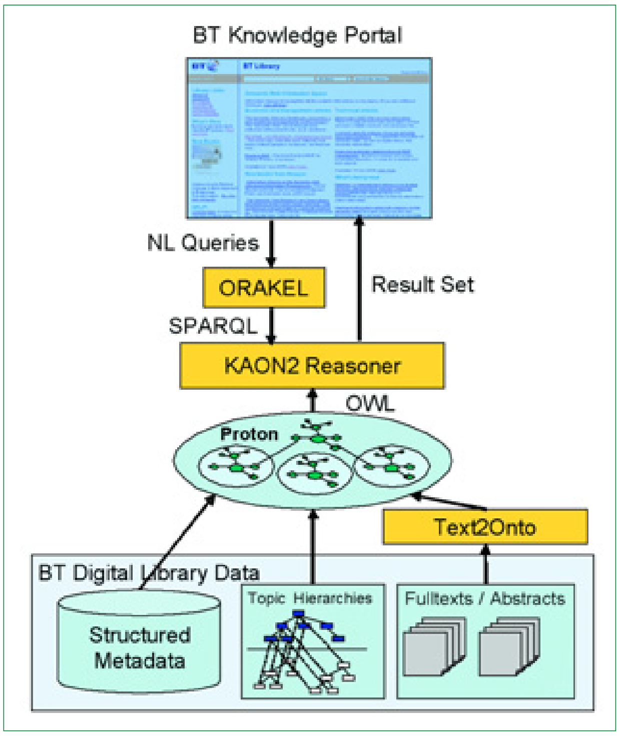

Figure 2: Digital Library Case Study

[3] http://www.sekt-project.com
[4] http://ontoware.org/projects/text2onto/
[5] http://jatke.opendfki.de/cgi-bin/trac.cgi

be plugged in, thus being very flexible and modular. To some extent this is an engineering problem but it also requires a clear understanding of the way ontology learning algorithms work.

On the other hand, most researchers have realized that the output of ontology learning algorithms is far from being perfect. As a consequence, to make the process controllable, we need an assessment of how certain an algorithm is in its predictions. Numeric confidence values of an algorithm in the certainty of a prediction could then be used as a basis to combine different algorithms compensating for the drawbacks and false predictions of each other. The representation of uncertainty and the combination of algorithms given their certainty are thus inherently coupled and represent one of the main open problems in the field of ontology learning. Though several proposals have recently emerged, the problem is far from being solved. In particular, coming up with a well-defined interpretation of certainty to give a sound basis for the combination of algorithms seems a non-trivial problem.

One of the most important challenges for ontology learning is the tight integration of automatic approaches with methodologies for ontology engineering and evaluation. Depending on the intended usage of an ontology, a sophisticated process of requirements specification, learning, evaluation and refinement may be necessary in order to create ontologies of sufficient conceptual preciseness. The user should be given the possibility to specify the objectives of the ontology learning process with respect to formal and subjective aspects (e.g. complexity or domain coverage), and to choose among a variety of modelling primitives suitable for his application. Whereas in some cases a bare taxonomy will be sufficient, other modelling tasks may require the use of more specialized ontological constructs or relationships such as causality or equivalence. At each stage of the ontology learning process or, at least, after each iteration of the whole engineering cycle, an automatic or manual evaluation of the learned ontology should take place in order to avoid the propagation of errors and to allow for a re-configuration of the algorithms by the user or the system itself. This evaluation must take into account the evidences or certainties generated by the individual ontology learning algorithms, the specification of the user requirements as well as formal constraints of the target ontology modelling language.

Conclusions and Outlook

In this article we have briefly discussed the state-of-the-art in ontology learning with respect to different ontology learning subtasks. Further, we have also discussed potential applications for ontology learning to support structuring information as well as in the field of text mining and information retrieval. We have also described how ontology learning algorithms have been applied within the SEKT project at British Telecom. Though the methods for ontology learning are still in their infancy, they have already today the potential to improve certain classical applications such as information retrieval and helping in structuring huge collections of resources by applying text mining techniques. We have also discussed challenges and trends within ontology learning research and highlighted two main trends: the trend to build flexible frameworks into which diverse algorithms can be plugged-in in a simple way, as well as the trend to extend algorithms towards predicting how confident they are in their predictions. The most obvious challenges in ontology learning research are on the one hand to come up with a suitable interpretation of what the confidences indicated by algorithms mean from a formal point of view, thus providing a basis on which to combine the results from different algorithms in a sound way. A further important question is how axiomatized the ontologies we learn can actually be and which methods we need to obtain ontologies for the whole bandwidth of potential applications.

Finally, coming back to the title of the article, it seems appropriate to provide an answer to the question whether the field of ontology learning is advanced enough to provide ontologies on demand. The answer to this question is actually twofold. We have seen on the one hand that for certain applications such as information retrieval and text mining, automatically learned ontologies already have the potential to provide an added value. These ontologies essentially provide an abstraction over plain words which Gruber refers to as `semi-formal´ ontologies. On the other hand, ontology learning techniques still seem to be in their infancy, since in many cases ontologies generated by off-the-shelf ontology learning methods do not meet the demands of the envisioned applications. Highly configurable methods and frameworks as well as a tight integration with manual or automatic ontology evaluation approaches (see for example [Guarino & Welty 2000]) will be required to ensure the applicability of ontology learning across different application areas. It seems crucial to invest in the development of new ontology engineering methodologies which are able to integrate the results of ontology learning systems into the ontology engineering process, keeping user input at a minimum while maximizing the quality of the ontologies with respect to a particular domain or application.

Acknowledgements

We acknowledge financial support from the projects SEKT and X-Media, funded by the European Union under grants EU-IST-IP-2003-506826 and EU-IST-FP6-026978, respectively, as well as the BMBF project SmartWeb, funded by the Germany Ministry of Education and Research.

Bibliography

[Agichtein & Gravano 2000] E. *Agichtein* and L. *Gravano*: Snowball: extracting relations from large plain-text collections. Proceedings of the ACM Conference on Digital Libraries, pp. 85-94, 2000.

[Baeza-Yates and Ribeiro-Neto 1999] R. *Baeza-Yates*, B. *Ribeiro-Neto*: Modern Information Retrieval. Addison Wesley, 1999.

[Berners-Lee et al. 2001] T. *Berners-Lee*, J. *Hendler*, O. *Lassila*: The Semantic Web. Scientific American 284(5), pp. 34-43, 2001.

[Bloehdorn et al. 2005] S. *Bloehdorn*, P. *Cimiano*, A. *Hotho*: Learning Ontologies to Improve Text Clustering and Classification. From Data and Information Analysis to Knowledge Engineering: Proceedings of the 29th Annual Conference of the German Classification Society (GfKl), 2005.

[Buitelaar & Cimiano 2006] P. *Buitelaar*, P. *Cimiano*: Tutorial Notes of the EACL Tutorial on Ontology Learning from Text, Trento, Italy, 2006.

[Brin 1998] S. *Brin*: Extracting Patterns and Relations from the World Wide Web. Proceedings of the WebDB Workshop at EDBT '98, pp. 172-183, 1998.

[Caraballo 1999] S.A. *Caraballo*: Automatic construction of a hypernym-labeled noun hierarchy from text. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 120-126, 1999.

[Cederberg & Widdows 2003] S. *Cederberg*, D. *Widdows*: Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. Proceedings of the International Conference on Natural Language Learning (CoNLL), pp. 111-118, 2003.

[Ciaramita et al. 2005] M. *Ciaramita*, A. *Gangemi*, E. *Ratsch*, J. *Saric*, I. *Rojas*: Unsupervised Learning of Semantic Relations between Concepts of a Molecular Biology Ontology. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp. 659-664, 2005.

[Cimiano et al. 2006] P. *Cimiano*, M. *Hartung*, E. *Ratsch*: Learning the Appropriate Generalization Level for Relations Extracted from the Genia Corpus. Proceedings of the International Conference on Language Resources and Evaluation (LREC), pp. 161-169, 2006.

[Cimiano & Völker 2005a] P. *Cimiano*, J. *Völker*: Towards large-scale, open-domain and ontology-based named entity classification. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), pp. 166-172, 2005.

[Cimiano & Völker 2005b] P. *Cimiano*, J. *Völker*: Text2Onto – A Framework for Ontology Learning and Data-driven Change Discovery. Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), volume 3513 of Lecture Notes in Computer Science, pp. 227-238, 2005.

[Cimiano and Staab 2005] P. *Cimiano*, S. *Staab*: Learning Concept Hierarchies from Text with a Guided Agglomerative Clustering Algorithm. Proceedings of the ICML 2005 Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods, 2005.

[Cimiano et al. 2005] P. *Cimiano*, A. *Hotho*, S. *Staab*: Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. Journal of Artificial Intelligence Research (JAIR), 24, pp. 305-339, 2005.

[Cimiano et al. 2004] P. *Cimiano*, S. *Handschuh*, S. *Staab*: Towards the self-annotating web. Proceedings of the 13th World Wide Web Conference, pp. 462-471, 2004.

[Davies et al. 2006] J. *Davies*, R. *Studer*, P. *Warren*: Semantic Web Technologies : Trends and Research in Ontology-based Systems. John Wiley & Sons, 2006.

[Etzioni et al. 2004] O. *Etzioni*, M.J. *Cafarella*, D. *Downey*, S. *Kok*, A.-M. *Popesu*, T. *Shaked*, S. *Soderland*, D.S. *Weld*, A. *Yates*: Web-scale information extraction in KnowItAll (preliminary results). Proceedings of the 13th World Wide Web Conference, pp. 100-110, 2004.

[Faure & Nedellec 1999] D. *Faure*, C. *Nedellec*: Knowledge Acquisition of Predicate Argument Structures from Technical Texts Using Machine Learning: The System ASIUM. Proceedings of the European Knowledge Acquisition Workshop (EKAW), pp. 329-334, 1999.

[Frantzi & Ananiadou 1999] K. *Frantzi*, S. *Ananiadou*: The C-value / NC-value domain independent method for multi-word term extraction. Journal of Natural Language Processing, 6(3), pp. 145-179, 1999.

[Gruber 1995] T.R. *Gruber*: Towards principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, Volume 43 , Issue 5-6, pp. 907-928, 1993.

[Gruber 2004] Interview for the Official Quarterly Bulletin of AIS Special Interest Group on Semantic Web and Information System 1(3), 2004.

[Guarino & Welty 2000] N. *Guarino*, C.A. *Welty*: A formal ontology of properties. Knowledge Acquisition, Modeling and Management, pp. 97-112, 2000.

[Holsapple & Joshi 2002] C.W. *Holsapple*, K.D. *Joshi*: A collaborative approach to ontology design. Communications of the ACM, 45(2), pp. 42-47, 2002.

[Mädche & Staab 2000] A. *Mädche*, S. *Staab*: Discovering Conceptual Relations from Text. Proceedings of the European Conference on Artificial Intelligence (ECAI), pp. 321-325, 2000.

[Manning & Schütze 1999] C.D. *Manning*, H. *Schütze*: Foundations of Statistical Natural Language Processing. MIT Press, 1999.

[Motik & Sattler 2006] B. *Motik*, U. *Sattler*: A Comparison of Techniques for Querying Large Description Logic Aboxes, accepted for publication at the 13th International Conference on Logic for Programming, Artificial Intelligence and Reasoning (LPAR), 2006.

[Pantel & Pennacchiotti 2006] P. *Pantel*, M. *Pennacchiotti*: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. Proceedings of the Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL), 2006.

[Paslaru Bontas et al. 2006] E. *Paslaru Bontas*, C. *Tempich*, Y. *Sure*: OntoCom: A Cost Estimation Model for Ontology Engineering. Proceedings of the 5th International Semantic Web Conference (ISWC), 2006.

[Pekar & Staab 2003] V. *Pekar*, S. *Staab*, Word classification based on combined measures of distributional and semantic similarity. Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), pp. 147-150, 2003.

[Pinto & Martins 2004] H.S. *Pinto*, J.P. *Martins*: Ontologies: How Can They Be Built? Knowledge and Information Systems, 6(4), pp. 441-464, 2004.

[Pinto et al. 2004] H.S. *Pinto*, S. *Staab*, C. *Tempich*: DILIGENT: Towards a fine-grained methodology for DIstributed, Loosely-controlled and evolvInG Engineering of oNTologies. Proceedings of the European Conference on Artificial Intelligence (ECAI), pp. 393-397, 2004.

[Schutz & Buitelaar 2005] A. *Schutz*, P. *Buitelaar*: RelExt: A Tool for Relation Extraction from Text in Ontology Extension. Proceedings of the 4th International Semantic Web Conference, (ISWC) pp. 593-606, 2005.

[Turney 2001] P.D. *Turney*: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. Proceedings of the 12th European Conference on Machine Learning (ECML), pp. 491-502, 2001.

[Völker et al. 2005] J. *Völker*, D. *Vrandecic*, Y. *Sure*: Automatic Evaluation of Ontologies (AEON). Proceedings of the 4th International Semantic Web Conference (ISWC), pp. 716-731, 2005.

Knowledge Representation, Research, Machine Learning, State-of-the-Art, Semantic Web, Text Mining

Lernen, maschinell, Semantisches Netz, Wissensrepräsentation, Forschung, Übersichtsbericht

Die Autoren

Dr. Philipp Cimiano

is a researcher and project leader at the Institute of Applied Informatics and Formal Description Methods (AIFB) at the University of Karlsruhe. He graduated in Computer Science at the University of Stuttgart in 2002. Philipp's main interests are in the fields of natural language processing, ontology learning, text mining and knowledge representation. Philipp is currently working on the European project X-Media as well as the German SmartWeb project. Together with Paul Buitelaar and Bernardo Magnini, he recently edited a collection of papers on ontology learning, appeared at IOS Press. He serves as a reviewer for international journals, conferences, and workshops.

cimiano@aifb.uni-karlsruhe.de
www.aifb.uni-karlsruhe.de/WBS/pci.

Johanna Völker

is a PhD student at the Institute of Applied Informatics and Formal Description Methods (AIFB) at the University of Karlsruhe. Her research interests include information extraction, text mining, ontology learning and the Semantic Web. She received a diploma in computer science with special focus on computational linguistics from Saarland University.

voelker@aifb.uni-karlsruhe.de
www.aifb.uni-karlsruhe.de/WBS/jvo.

Prof. Dr. Rudi Studer

is the head of the knowledge management research group at AIFB, and obtained a Diploma in Computer Science at the University of Stuttgart in 1975. In 1982 he was awarded a Doctor degree in Mathematics and Computer Science (Dr. rer. nat.) at the University of Stuttgart, and 1985 he obtained his Habilitation in Computer Science. From July 1985 to October 1989 he was project leader and manager at IBM Germany, Institute of Knowledge Based Systems. Since November 1989 he has been a full professor in Applied Computer Science at the Institute AIFB at the University of Karlsruhe. Since then, he led his research group to become one of the world leading institutions in Semantic Web technology, and he played a leading role in establishing highly acknowledged international conferences and journals in this area. He is president of the Semantic Web Science Association, Technical Director of the EU-funded SEKT project, joint Research Area Manager of the EU-funded Knowledge-Web network, and a member of numerous programme committees and editorial boards, including position of joint editor-in-chief of the Journal on Web Semantics. His current research interests span over the main topics important for Semantic Web technology, including knowledge management, knowledge engineering, intelligent web brokers, and knowledge discovery and learning.

studer@aifb.uni-karlsruhe.de
www.aifb.uni-karlsruhe.de/WBS/rst.