Chris Biemann

Ontology Learning from Text: A Survey of Methods

1 Introduction

After the vision of the Semantic Web was broadcasted at the turn of the millennium, *ontology* became a synonym for the solution to many problems concerning the fact that computers do not understand human language: if there were an ontology and every document were marked up with it and we had agents that would understand the mark-up, then computers would finally be able to process our queries in a really sophisticated way. Some years later, the success of Google shows us that the vision has not come true, being hampered by the incredible amount of extra work required for the intellectual encoding of semantic mark-up – as compared to simply uploading an HTML page. To alleviate this acquisition bottleneck, the field of ontology learning has since emerged as an important sub-field of ontology engineering.

It is widely accepted that ontologies can facilitate text understanding and automatic processing of textual resources. Moving from words to concepts not only mitigates data sparseness issues, but also promises appealing solutions to polysemy and homonymy by finding non-ambiguous concepts that may map to various realizations in – possibly ambiguous – words.

Numerous applications using lexical-semantic databases like WordNet (Miller, 1990) and its non-English counterparts, e.g. EuroWordNet (Vossen, 1997) or CoreNet (Choi and Bae, 2004) demonstrate the utility of semantic resources for natural language processing.

Learning semantic resources from text instead of manually creating them might be dangerous in terms of correctness, but has undeniable advantages: Creating resources for text processing from the texts to be processed will fit the semantic component neatly and directly to them, which will never be possible with general-purpose resources. Further, the cost per entry is greatly reduced, giving rise to much larger resources than an advocate of a manual approach could ever afford. On the other hand, none of the methods used today are good enough for creating semantic resources of any kind in a completely unsupervised fashion, albeit automatic methods can facilitate manual construction to a large extent.

The term *ontology* is understood in a variety of ways and has been used in philosophy for many centuries. In contrast, the notion of ontology in the field of computer science is younger – but almost used as inconsistently, when it comes to the details of the definition.

The intention of this essay is to give an overview of different methods that learn ontologies or ontology-like structures from unstructured text. Ontology learning from other sources, issues in description languages, ontology editors, ontology merging and ontology evolving transcend the scope of this article. Surveys on ontology learning from text and other sources can be found in Ding and Foo (2002) and Gómez-Pérez

and Manzano-Macho (2003), for a survey of ontology learning from the Semantic Web perspective the reader is referred to Omelayenko (2001).

Another goal of this essay is to clarify the notion of the term *ontology* not by defining it once and for all, but to illustrate the correspondences and differences of its usage.

In the remainder of this section, the usage of *ontology* is illustrated very briefly in the field of philosophy as contrasted to computer science, where different types of ontologies can be identified.

In section 2, a variety of methods for learning ontologies from unstructured text sources are classified and explained on a conceptual level. Section 3 deals with the evaluation of automatically generated ontologies and section 4 concludes.

1.1 Ontology in philosophy

In philosophy, the term *ontology* refers to the study of existence. In this sense, the subject is already a central topic of *Aristotle's Categories* and in all metaphysics. The term was introduced in the later Renaissance period, see Ritter and Gründer (1995), as "*lat.* philosophia de ente". In the course of centuries, *ontology* was specified in different ways and covered various aspects of metaphysics. It was sometimes even used as a synonym for this field. Further, the distinction between ontology and theology was not at all times clear and began to emerge in the 17th century.

For Leibniz, the subject of *ontology* is everything that can be recognized (*germ.* erkannt). Recognition (*germ.* Erkenntnis) as a basis of metaphysics is criticised by Kant, who restricts ontology to a propaedeutical element of metaphysics, containing the conditions and the most fundamental elements of all our recognition (*germ.* Erkenntniß) a priori.

The relation of ontology to logic was introduced by Hegel and later strengthened by Husserl, who defends the objectivity of logical entities against subjectivation and replaces the notion of logical terms as psychical constructions with "ideal units" that exist a priori. Ontology in this context can be divided into two kinds: *formal ontology* that constitutes itself as a theory of all possible forms of theories, serving as science of sciences, and *regional* or *material ontologies* that are the a priori foundations of empirical sciences (Husserl, 1975). The latter notion paved the way to *domain-specific ontologies*, see section 1.2.

For computer science, the most influential definition has been given by Quine (cf. Quine, 1969), who binds scientific theories to ontologies. As long as a theory holds (because it is fruitful), theoreticians perform an ontological commitment by accepting the a priori existence of objects necessary to prove it. A consequence of his famous quote "to be is to be the value of a bound variable" is: As long as scope and domain of quantified variables (objects) are not defined explicitly by an ontology, the meaning of a theory is fuzzy. Ontologies in the sense of Quine are the outcome of empirical theories, and hence they also need to be justified empirically.

To subsume, ontology abstracts from the observable objects in the world and deals with underlying principles of existence as such.

1.2 Ontologies in Computer Science

Ontology in computer science is understood not as general as in philosophy, because the perception of ontologies is influenced by application-based thinking. But still ontologies in computer science aim at explaining the world(s), however, instead of embracing the whole picture, they only focus on what is called a *domain*. A domain is, so to speak, the world as perceived by an application. Example: The application of a fridge is to keep its interior cold and that is reached by a cooling mechanism which is triggered by a thermostat. So the domain of the fridge consists only of the mechanism and the thermostat, not of the food in the fridge, and can be expressed formally in a fridge ontology. Whenever the application of the fridge is extended, e.g. to illuminate the interior when the door is opened, the fridge ontology has to be changed to meet the new requirements. So much about the fridge world. In real applications, domains are much more complicated and cannot be overseen at a glance.

Ontologies in computer science are specifications of shared conceptualizations of a domain of interest that are shared by a group of people. Mostly, they build upon a hierarchical backbone and can be separated into two levels: upper ontologies and domain ontologies.

Upper ontologies (or foundation ontologies), which describe the most general entities, contain very generic specifications and serve as a foundation for specializations. Two well-known upper ontologies are SUMO (Pease and Niles, 2002) and CyC (Lenat, 1995). Typical entries in upper ontologies are e.g. "entity", "object" and "situation", which subsume a large number of more specific concepts. Learning these upper levels of ontologies from text seems a very tedious, if not impossible task: The connections as expressed by upper ontologies consist of general world knowledge that is rather not acquired by language and is not explicitly lexicalized in texts.

Domain ontologies, on the other hand, aim at describing a subject domain. Entities and relations of a specific domain are sometimes expressed directly in the texts belonging to it and can eventually be extracted. In this case, two facts are advantageous for learning the ontological structures from text: The more specialized the domain, the less is the influence of word sense ambiguity according to the "one sense per domain"-assumption in analogy to the "one sense per discourse"-assumption (Gale et al., 1993). Additionally, the less common-knowledge a fact is, the more likely it is to be mentioned in textual form.

In the following section, distinctions between different kinds of ontologies and other ways of categorizing the world are drawn.

1.3 Types of Ontologies

John Sowa (Sowa, 2003) classifies ontologies into three kinds. A *formal ontology* is a conceptualization whose categories are distinguished by axioms and definitions. They are stated in logic that can support complex inferences and computations. The knowledge representation community defines ontology in accordance as follows:

> "[An ontology is] a formal, explicit specification of a shared conceptualization. 'Conceptualization' refers to an abstract model of phenomena in the world by having identified the relevant concepts of those phenomena. 'Explicit' means that the type of concepts used, and the constraints on their use are explicitly defined. 'Formal' refers to the fact that the ontology should be machine readable. 'Shared' reflects that ontology should capture consensual knowledge accepted by the communities." (Gruber, 1993; Ding and Foo, 2002)

As opposed to this, categories in *prototype-based ontologies* are distinguished by typical instances or prototypes rather than by axioms and definitions in logic. Categories are formed by collecting instances extensionally rather than describing the set of all possible instances in an intensional way, and selecting the most typical members for description. For their selection, a similarity metric on instance terms has to be defined.

The third kind of ontology are *terminological ontologies* that are partially specified by subtype-supertype relations and describe concepts by concept labels or synonyms rather than prototypical instances, but lack an axiomatic grounding. A well known example for a terminological ontology is WordNet (Miller, 1990).

Figure (1) illustrates different ontology paradigms for a toy example food domain divided into vegetarian and non-vegetarian meals.

All of these paradigms have their strengths and weaknesses. Formal ontologies directly induce an inference mechanism. Thus, properties of entities can be derived when needed. A drawback is the high effort of encoding and the danger of running into inconsistencies. Further, exact interference may become intractable in large formal ontologies.

Terminological and prototype-based ontologies cannot be used in a straightforward way for inference, but are easier to construct and to maintain. A disadvantage of the prototype-based version is the absence of concept labels, which makes it impossible to answer queries like "Tell me kinds of cheese!". Due to the absent labeling during construction, they are directly induced by term clustering and therefore easier to construct but less utilizable than their terminological counterparts.

A distinction that causes confusion are the notions of taxonomy versus ontology, which are occasionally used in an interchangeable way. Taxonomies are collections of entities ordered by a classification scheme and usually arranged hierarchically. There is only one type of relation between entries, mostly the IS-A or PART-OF relation. This corresponds to the notion of terminological ontologies. For formal ontologies, the concepts together with IS-A relations form the taxonomic backbone of the ontology.

Another kind of resource which is a stepping stone towards ontologies are thesauri like Roget's Thesaurus (Roget, 1852) for English or Dornseiff (Dornseiff, 2004) for German. A thesaurus contains sets of related terms and thus resembles a prototype-based ontology. However, different relations are mixed: a thesaurus contains hierarchy relations amongst others, but they are not marked as such.

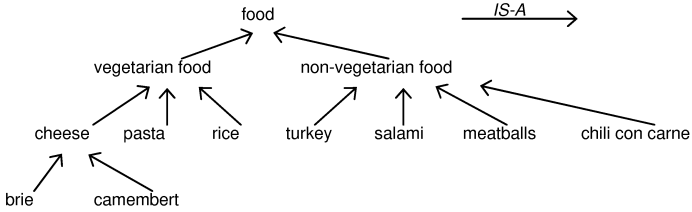
Formal ontology

Axioms:
food(brie), food(camembert), food(turkey), food(meatballs), food(chili con carne), meat(turkey), meat(minced meat), part_of(minced meat, chili con carne), part_of(minced meat, meatballs)

veg_food(x) = { x | food(x) $\wedge$ ($\neg$part_of(y,x) $\wedge$ meat(y)) $\wedge$ $\neg$meat(x) }
non_veg_food(x) = { x | food(x) $\wedge$ ((part_of(y,x) $\wedge$ meat(y)) $\vee$ meat(x)) }

Possible to derive: "turkey" and "chili con carne" are non-vegetarian foods

Terminological ontology



Prototype-based ontology

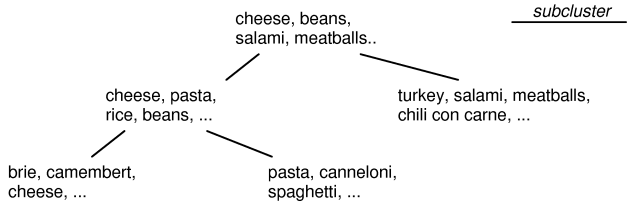


Figure 1: Formal vs. terminological vs. prototype-based food ontology.

2 Learning Ontologies from unstructured Text

Ontologies can be learnt from various sources, be it databases, structured and unstructured documents or even existing preliminaries like dictionaries, taxonomies and directories. Here, the focus is on acquisition of ontologies from unstructured text, a format that scores highest on availability but lowest on accessibility.

Most approaches use only nouns as the bricks for ontology building and disregard any ontological relations between other word classes.

To a large extent, the methods aim at constructing IS-A-related concept hierarchies rather than full-fledged formal ontologies. Other subtype-supertype relations like PART-OF are examined much less.

One underlying assumption for learning semantic properties of words from unstructured text data is Harris' distributional hypothesis (Harris, 1968), stating that similar words tend to occur in similar contexts. It gives rise to the calculation of paradigmatic relations (cf. Heyer et al., 2005), called 'associations' in de Saussure (1916). We shall see that the notion of context as well as the similarity metrics differs considerably amongst the approaches presented here.

Another important clue is the use of patterns that explicitly grasp a certain relation between words. After the author who introduced patterns such as "*X*, *Y*s and other *Z*s" or "*W*s such as *X*, *Y* and *Z*", they are often referred to as Hearst-patterns (Hearst, 1992), originally used to extract IS-A relations from an encyclopedia for the purpose of extending WordNet. Berland and Charniak (1999) use similar kinds of patterns to find instances of the PART-OF relation.

As learning from text usually involves statistics and a corpus, using the world wide web either as additional resource or as the main source of information is often a possibility to avoid data sparseness as discussed in Keller et al. (2002) and carried out e.g. by Agirre et al. (2000) and Cimiano and Staab (2004).

Ontology learning techniques can be divided in constructing ontologies from scratch and extending existent ontologies. The former comprises mostly clustering methods that will be described in section 2.1, the latter is a classification task and will be treated in section 2.2. Approximately, this is the distinction between unsupervised versus supervised methods, although we shall see that some clustering approaches involve supervision in intermediate steps.

Section 2.3 summarizes research undertaken in semantic lexicon construction, which is a related task to ontology learning, although the representation of results might differ. In section 2.4, the view of ontology learning as an Information Extraction exercise is discussed.

2.1 Clustering for Ontology Learning

In hierarchical clustering, sets of terms are organized in a hierarchy that can be transformed directly into a prototype-based ontology. For clustering, a distance measure on terms has to be defined that serves as the criterion for merging terms or clusters of terms. The same measure can be used – if desired – to compute the most typical instances of a concept as the ones closest to the centroid (the hypothetical 'average' instance of a set). Crucial to the success of this methodology is the selection of an appropriate measure of semantic distance and a suitable clustering algorithm.

An overview of clustering methods for obtaining ontologies from different sources including free text can be found in Maedche and Staab (2004). In principle, all kinds of clustering methods – be it agglomerative or divisive – can be applied to all kinds of representations, be it vector space (Salton et al., 1975), associative networks (Heyer and Witschel, 2005) or set-theoretic approaches as presented in Cimiano et al. (2004). Here, the focus will be on just a few, illustrative methods.

Methods based on distributional similarity Methods using distributional similarity can be divided into syntactic and window-based approaches.

Syntactic approaches make use of similarity regarding predicate-argument relations (i.e. verb-subject and verb-object relations), the usage of adjective modifiers or subjective predicates is rare.

An early paper on semantic clustering is Hindle (1990), which aims at finding semantically similar nouns by comparing their behavior with respect to predicate-argument structures. For each verb-subject and verb-object pair in his parsed 6 million word corpus, he calculates co-occurrence weights as the mutual information within the pairs. Verb-wise similarity of two nouns is the minimum shared weight, and the similarity of two nouns is the sum of all verb-wise similarities. The exemplified analysis of this similarity measure exhibits mostly homogeneous clusters of nouns that act or are used in a common way.

For obtaining noun hierarchies from text, Pereira et al. (1993) chose an encyclopedia as a well-suited textual resource for a divisive clustering approach based on verb-object relations, allowing the nouns to be members in multiple clusters.

A whole class of syntactic approaches is subsumed in the Mo'K workbench (Bisson et al., 2000), which provides a framework to define hierarchical term clustering methods based on similarity in contexts limited to specific syntactic constructions. In the same work, comparative studies between different variants of this class are presented, including ASIUM (Faure and Nédellec, 1998; Dagan et al., 1994). Another paper on using selectional preferences is e.g. Wagner (2000).

A different direction is using methods that produce paradigmatic relations as candidate extraction mechanism without syntactic pre-processing. A well-known source of paradigmatic relations is the calculation of second-order co-occurrences, which does not rely on parsing. While (first-order) co-occurrences rate pairs of word high that occur together often in a certain text window, second order co-occurrences are words that have similar distributions of first-order co-occurrences (see e.g. Ruge (1992), Schütze (1998), Rapp (2002), Biemann et al. (2004) – this corresponds roughly to Rieger's δ-abstraction (Rieger, 1981; Leopold, 2005)). The context definition of these methods is mostly not restricted to any syntactic construction, which introduces more noise but keeps the method language-independent. It can be argued that given a sufficient corpus size, equal results to syntactically aided methods might be achieved, see e.g. Pantel et al. (2004). However, as the underlying bag-of-words simplification of window-based methods abstracts from the order of the words, no clues for the relation between candidate pairs can be drawn directly from this data, making these approaches on their own not viable for the construction of ontologies from scratch.

While there does not seem to be an alternative to use patterns in order to alleviate the labeling problem, the action of naming super-concepts is not necessary when aiming at a prototypical ontology, such as in Paaß et al. (2004): here, a hierarchical extension to Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) is introduced. PLSA (like LSA – (cf. Deerwester et al., 1990)) assumes latent concepts, which are playing the role of an intermediate concept layer: the probability of seeing a word w in a document d is

the sum of the product probabilities of d belonging to concepts c and w being generated when c is present. To introduce hierarchical dependencies, the probability mass is split between sub- and super-concepts. In an experiment, a fixed 4-level hierarchy with 1, 14, 28 and 56 nodes on the levels was defined. The words with the highest probability per concept constitute the entries of the prototypical ontology. While results look impressive, a clear drawback is the predefined structure of the hierarchy.

Methods based on extraction patterns The other possibility is to use explicit clues, like Hearst-patterns.

Caraballo (1999) constructs a terminological ontology from text in the following way: noun candidates from a newspaper corpus are obtained by considering conjunction and appositive data. For all nouns, a co-occurrence matrix is set up. Similarity between two nouns is calculated by computing the cosine between their respective vectors and used for hierarchical bottom-up clustering. For labelling this hierarchy in a post-processing step, Hearst-patterns are used for finding hypernym candidates, which are placed as common parent nodes for clusters, if appropriate. Evaluated by human judgement, the method performs at about 35-55% precision.

A similar approach is presented by Cimiano and Staab (2005) who also cluster nouns based on distributional similarity and use Hearst-patterns, WordNet and patterns on the web as a hypernym oracle for constructing a hierarchy. Unlike as in Caraballo (1999), the hypernym sources are directly integrated into the clustering, deciding for each pair of nouns how they should be arranged into the hierarchy. The resulting taxonomy outperforms Caraballo's when evaluating the outcome against a reference ontology (see section 3).

2.2 OL as a classification task

Given an existing ontology, its extension can be viewed as a classification task: features of the existing data are used as a training set for Machine Learning, which produces a classifier for previously unknown instances.

One possibility is to utilize the hierarchical structure in a decision tree, as proposed in Alfonseca and Manandhar (2002). When inserting new concepts, it is tested whether they fit best to the actual node or one of the daughter nodes. The tree is traversed top-down from the root until an appropriate position is found. The largest problem here is the general nature of top-level concepts that leads to taking the wrong path in the beginning of the process, which can be alleviated by propagating the signatures of lower-level concepts one step upwards. For around 1200 concepts, an accuracy of about 28% is reported. A related approach is Witschel (2005), which substitutes the syntactic dependencies for similarity by comparing words only on sentence-based co-occurrences. A small sub-tree of an existing WordNet-like hierarchy is used as training and test data. Propagating the semantic descriptions iteratively upwards to the root, the approach is biased towards putting new words into larger sub-trees. While Witschel's results are better, this might be due to the smaller number of concept classes.

In Fleischman and Hovy (2002), only eight categories for named entities denoting persons are considered. They examine five machine learning approaches on features based on preceding and following word N-grams which are combined into concepts using WordNet, reporting 70% accuracy.

Placing words into WordNet where the concentration of words with similar distributional characteristics is highest is conducted by Widdows (2003). He arrives at about 80% precision for common nouns, 34% for proper nouns and 65% for verbs.

How to enlarge WordNet by assigning appropriate named entities to the leaves using the Google index is discussed in Paşca (2005).

2.3 Ontology Learning as Semantic Lexicon Construction

The similarities between the construction of semantic lexicons and lexical ontologies – be it terminological or prototype-based – are striking. Both encode semantic similarities between terms and they abstract terms to concepts. Whereas semantic lexicons often attach semantic categories to words and do not structure the set of words internally any further (although semantic lexicons like e.g. HaGenLex (Helbig, 2001) are organized in a flat hierarchy of categories), ontologies aim at explaining all possible relations between concepts, being more fine-grained. Nevertheless, words with the same semantic label should be found in the same region of the ontology, which makes part of the methodology for automatic construction applicable to both tasks.

Let us assume that we have a small semantic lexicon, given by a set of categories, which are formed each by a set of words. Using a text corpus, we want to extend this lexicon.

Mostly, bootstrapping approaches have been used to tackle this problem. On the one hand, because bootstrapping can iteratively use previously learnt examples, which reduces the minimal size of the seed lexicon. On the other hand it does not necessarily need negative examples for learning, making the procedure viable for learning single semantic categories. The largest problem that bootstrapping methods have to face is error-propagation: misclassified items will lead to the acquisition of even more misclassified items. Various attempts have been made to minimize this thread.

In general, bootstrapping starts with a small set of seeds as current category. For every candidate item, the similarity to the current category is computed and the most similar candidates are added to the current category. These steps are conducted iteratively until a stopping criterion holds; sometimes the process is deliberately stopped after about 50-200 iterations.

Riloff and Shepherd (1997) were the first to apply bootstrapping for building semantic lexicons, extending one category at a time. Their context definition is one noun to the left and one noun to the right for head nouns in sentences. Collecting the contexts of the current category set, they calculate a score for each word by checking the relative frequency of the word appearing in the category's contexts. Amongst the first 100 words retrieved by the algorithm for categories of a seed size around 50, about 25% were judged correct by human decision. In Riloff and Jones (1999) not only classes are

assigned to words, but also the confidence of contexts supporting a class is estimated. Contexts in this work are patterns such as ''headquartered in <x>'' or ''to occupy <x>''. Moreover, only the top 5 candidates are added to the knowledge base per step, alleviating error-propagation to a precision of about 46%-76% after 50 iterations. Further improvement was gained in Thelen and Riloff (2002), where multiple categories are learned at the same time to avoid too large single categories consisting of a mixture with several other categories. In that way, about 80% accuracy for the first couple of hundred new words can be reached. This complies well with the structuralist notion of semantics being defined in a negative way (de Saussure, 1916; Eco, 1977): A category "grows" in the space of meaning as long as it meets the border of another category.

Building multiple categories simultaneously is also used in Biemann and Osswald (2005), who extend a semantic lexicon for the use of semantic parsing. As contexts, only modifying adjectives of nouns are taken into account. Semantic classes of known nouns are inherited via the modifying adjectives to previously unclassified nouns. In experiments using a co-occurrence significance measure to consider merely typical modifiers, the authors report to succeed in doubling their resource of 6000 nouns in 50 categories with an accuracy of about 80%.

As opposed to these shallow approaches, Roark and Charniak (1998) look for words occurring together in syntactical formations that involve full parsing of the corpus. A radical break with syntactical pre-processing is conducted in Biemann et al. (2004), where a lexical-semantic resource is extended without using any tagging or parsing, merely by using sentence-based co-occurrence statistics. A word is added to a category if many words of the category occur with it within a sentence window. While scores are differing strongly for selected categories, the approach serves as a language-independent baseline.

2.4 Information Extraction for Ontology Population

In Information Extraction (IE, see Grishman (1997) for a survey), templates containing roles, relations, temporal and time information to describe possible situations are encoded. The task is to fill the templates' slots by extracting relevant data from documents. IE proceeds in a situative way: instantiated templates are attached to the texts from which they have been extracted. Ontologies, on the other hand, encode conceptualizations that are not bound to specific text snippets but apply in general. Nevertheless, templates can be defined in IE systems like GATE (Bontcheva et al., 2004) and the standard IE extraction mechanisms can be employed to fill these templates, producing eventually more powerful and flexible extraction rules than the patterns mentioned before.

IE systems are historically suited for the extraction of named entities. This is why they are mainly used to find instances of concepts (like chocolate companies) and relations (like employer – employee) rather than the concepts themselves: they can be better used for populating than for constructing ontologies. After collecting all the situative template instantiations, pruning has to be applied to keep only relations that occur frequently and with high confidence.

In Brin (1998), the DIPRE system is laid out that bootstraps the AUTHOR-OF relation between writers and book titles by automatically creating extraction patterns that heavily rely on HTML-tags, but also use clues from unformatted text. Using a DIPRE-like architecture, the SNOWBALL system (Agichtein and Gravano, 2000) learns patterns for free text that has been tagged by a named entity recognizer and uses them to extract instances similar to a few user-provided example tuples, never attempting to extract all the information from each document. For example, the SNOWBALL-pattern ''<LOCATION>-based <ORGANISATION>'' extracts headquarters of companies with high precision. Sufficient recall is ensured by using a large corpus.

3 Evaluation

As ontology learning just emerged recently as a field of its own, there are not many gold standards that could be used for evaluation. Further, the desired result of ontology learning is not a simple list with binary classifications, but a far more complicated structure. To make it even worse, there is "no clear set of knowledge-to-be-acquired" (Brewster et al., 2004), not even for very specialized domains. As Smith (2004) claims, there are several possibilities of conceptualizations for one domain that might differ in their usefulness for different groups of people, but not in their soundness and justification. So even if the outcome of an algorithm does not compare well with a manually built ontology, how can its quality be judged?

Of course, there is always the option of manual evaluation, with its well-known drawbacks of being subjective and time-consuming. For complicated tasks like ontology learning, a comparably low inter-annotator agreement can be expected, which in turn means that several annotators have to judge the results to arrive at consistent figures.

But maybe it is not the ontology itself that is in the focus of interest, but its application. Learning ontologies is a goal of its own, but ontologies are usually just a resource that should improve performance on NLP tasks. Measuring improvements of ontology-supported approaches depicts best the gain for the application in focus, but it unfortunately does not provide direct scores for the ontology itself.

In the remainder of this section, several possibilities to conduct an automatic evaluation on ontologies are discussed.

3.1 Evaluation against a Gold Standard

The question on how to compare two taxonomies or ontologies is first dealt with in Maedche and Staab (2002), who show ways to compare them on lexical and on conceptual level. For the lexical level, they measure the lexical overlap between the concept names in a variant-robust way. For comparing the taxonomic backbones of two ontologies, the notion of semantic cotopy is introduced. Semantic cotopy of a concept is the union of all its sub- and super-concepts, approximating its intensional semantics. The averaged taxonomical similarity is determined by the maximal overlap of semantic cotopies. Further, the authors provide ways to compare the relations of two ontologies

and evaluate their measures by an empirical study, using the tourism ontology developed within the GETESS project (Staab et al., 1999).

When aiming at taxonomy relations, it is possible to compare results of an algorithm with lexical-semantic nets like WordNet, as employed by e.g. Wagner (2000) and Witschel (2005). Yet, whenever a relation is not found in the gold standard, the algorithm might be wrong or the gold standard might be incomplete. This even holds for large resources – Roark and Charniak (1998) report that 60% of the terms generated by their semantic class learner could not be found in WordNet.

In Brewster et al. (2004) a comparison of ontologies with automatically extracted keywords from text corpora is proposed. The method measures lexical overlap as a score of how much the ontology fits the texts, but merely in a bag-of-words fashion, disregarding internal structure.

3.2 Application-based Evaluation

Recent years saw an increasing amount of research using WordNet to improve any kind of NLP application. The bulk of these applications can in turn be used for evaluating automatically created semantic resources. In the following paragraphs, setups for an application-based evaluation of ontologies are discussed.

Document clustering and classification Document similarity is usually measured by comparison of document vectors in a vector space (Salton et al., 1975), where each dimension in the space represents one term. Ambiguity and variability of natural language might cause several concepts to be mapped onto one dimension and several dimensions to be used for one concept, resulting in spurious similarity scores. This is the main motivation to use LSA (Deerwester et al., 1990), which reduces the number of dimensions by just considering main components as determined by singular value decomposition. But LSA has a number of drawbacks, including bad scalability and black-box-like latent concepts. With a domain-specific ontology, terms that are found in or around the same concept can be mapped into one dimension. On the other hand, terms that are present in many concepts due to their semantic ambiguity can be excluded or disambiguated, see next paragraph.

The clustering induced by the similarity measure can be compared to pre-categorized collections such as the Reuters corpus (Reuters Corpus, 2000). It is also possible to train a classifier and compare its performance between presence and absence of ontology information. Evaluation using document similarity will favor ontologies that keep similar terms in similar places, possibly in a flat and not very fine-grained hierarchy.

In Heinrich et al. (2005), two latent concept methods for constructing a prototype-based ontology are compared by measuring their effects on document clustering. As latent methods include the notion of a document into their models and can be applied to cluster words as well as documents, the choice seems natural. The ontology is used for dimensionality reduction in document clustering, which is compared to a gold standard.

Word sense disambiguation The task of word sense disambiguation (WSD) is to choose the appropriate sense for ambiguous words from a predefined inventory of senses. For English, WSD methods are usually evaluated on the SENSEVAL corpora (Kilgarriff, 1998), using WordNet as sense dictionary. Senses are assigned according to the ambiguous words' contexts: Either contexts are compared to glosses and terms close to the different concepts in WordNet (unsupervised WSD) or to context profiles per sense acquired from a training corpus (supervised WSD). Using WSD for the evaluation will favour ontologies that distinguish well between the different senses of words. WSD was successfully supported by semantic resources obtained from large corpora by Gliozzo et al. (2005), where terms are mapped to domains using LSA with a large number of dimensions.

Information Retrieval and Question Answering After various attempts to use query expansion methods in order to provide better coverage for information retrieval (see e.g. Ruge (1992); Stamou and Christodoulakis (2005)), this direction to improve information retrieval has been largely abandoned as it usually decreases precision too much without considerably improving recall. A possible reason is that users actually look for what they have been typing in and not for hypernyms or even synonyms. Another problem is the lack of disambiguation clues in the query which causes the query expansion mechanism to over-generate even worse.

But ontologies can be used in other parts of the retrieval process. Taxonomies that are build automatically from web data are used by Sánchez and Moreno (2005) to group query results returned by a search engine. In this case, the user's behavior of accepting or rejecting the interface is the instance of judgement. Improving question answering by overcoming the shortfalls of the bag-of-words model is the objective of e.g. Leveling and Hartrumpf (2005). Here, a semantic lexicon forms the background knowledge for semantic parsing, which yields a semantic representation much more precise than simply considering presence or absence of terms. Extending the lexicon as described in section 2.3 should result in higher performance.

Using Information Retrieval and Question Answering tasks for evaluation will promote ontologies with high coverage, as these applications are usually tested in a generic rather than in a domain-specific setting.

Co-reference Resolution The goal of co-reference resolution is to detect words that form a referent chain in a text. These chains mostly consist of pronouns, but also synonyms, hypernyms and part-whole related terms can refer to a previously mentioned entity. Co-reference resolution can be viewed as the classification task of finding the right antecedent for a referent using e.g. grammatical, contextual and morphological features. The evaluation framework for English co-reference resolution, which is not an application itself but rather a pre-processing step for methods like summarization, abstracting and information extraction, are the MUC-6 and MUC-7 corpora (Chinchor, 1998). The use of semantic resources, however, is scarcely encountered for co-reference or anaphora resolution. An exception is Hoste (2005), where WordNet and the Dutch

part of EuroWordNet are used for additional features, which bring about only a small gain because of lacking coverage. At first glance, it seems that ontologies can only support co-reference resolution in the rare cases of nouns referring to other nouns that are semantically related, but not in the default case of pronouns referring back to noun phrases. But there is the possibility of using the semantic role of the pronoun to find antecedents that are compatible, e.g. as subject or object of the pronoun's sentence's predicate, as pointed out by Johansson et al. (2005). As there is plenty of room for improvement in co-reference and anaphora resolution, this might be a suitable task to evaluate ontologies that encode semantic roles additionally to hierarchical relations.

4 Conclusion

After clarifying the usage of the term *ontology*, a variety of methods have been described how to construct and extend ontologies using unstructured text sources. We have then been looking at approaches that are directly labeled with *ontology learning*, complemented by a consideration of earlier work that has similar goals despite differing terminology. Further, various scenarios for ontology evaluation have been conveyed.

Currently, ontology learning cannot fulfill the promises that its name suggests. As far as prototype-based ontologies are concerned, clustering might yield more or less semantically coherent sets of words, but will not be of great help for carrying out the crucial step from terms to concepts. Taxonomical ontologies can be learnt as far as the relations are explicitly mentioned in the text and extractable by patterns that are scarcely met in real life texts. For circumventing the problem of possible pattern mismatches (i.e. "life is a highway") even more text has to be considered, resulting in very small taxonomies as opposed to the size of the corpus, as pointed out by Brewster et al. (2005).

Especially when comparing the requirements for formal ontologies as formulated by the Semantic Web community and the structures learnable from text as described, one has to state that the 'self-annotating web' will remain a vision for a long time.

But maybe the task is ill-defined. It is beyond doubt that modeling semantics will carry natural language processing further, as it has reached a state where further improvement of systems would in fact need rather more language understanding than more rules or more training examples. It is an open question, however, whether formal specifications are the only way to reach the goal, or whether the manual approach of hand-coding semantics will be outperformed by inconsistent, statistical black-box methods again.

5 Acknowledgements

The author would like to thank Gerhard Heyer and Christer Johansson for useful comments. This work was partially carried out at MULTILINGUA, University of Bergen, supported by the European Commission under the Marie Curie actions.

References

Agichtein, E. and L. Gravano (2000). Snowball: Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM International Conference on Digital Libraries*.

Agirre, E., O. Ansa, E. Hovy, and D. Martinez (2000). Enriching very large ontologies using the WWW. In *Proceedings of the ECAI 2000 Workshop on Ontology Learning*, Berlin, Germany.

Alfonseca, E. and S. Manandhar (2002). Extending a lexical ontology by a combination of distributional semantics signatures. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2002)*, Berlin, pp. 1–7. Springer.

Berland, M. and E. Charniak (1999). Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*.

Biemann, C., S. Bordag, and U. Quasthoff (2004). Automatic acquisition of paradigmatic relations using iterated co-occurrences. In *Proceedings of LREC 2004*, Lisboa, Portugal.

Biemann, C. and R. Osswald (2005). Automatische Erweiterung eines semantikbasierten Lexikons durch Bootstrapping auf großen Korpora. In B. Fisseni, H.-C. Schmitz, B. Schröder, and P. Wagner (Eds.), *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. Beiträge zur GLDV-Tagung 2005, Universität Bonn*, Frankfurt am Main. Peter Lang.

Biemann, C., S.-I. Shin, and K.-S. Choi (2004). Semiautomatic extension of CoreNet using a bootstrapping mechanism on corpus-based co-occurrences. In *Proceedings of the 20th Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, pp. 1227–1232.

Bisson, G., C. Nédellec, and L. Cañamero (2000). Designing clustering methods for ontology building – the Mo'K workbench. In *Proceedings of the ECAI 2000 Workshop on Ontology Learning*, Berlin, Germany.

Bontcheva, K., V. Tablan, D. Maynard, and H. Cunningham (2004). Evolving GATE to meet new challenges in language engineering. *Natural Language Engineering 10*(3/4), 349–373.

Brewster, C., H. Alani, S. Dasmahapatra, and Y. Wilks (2004). Data driven ontology evaluation. In *Proceedings of LREC 2004*, Lisboa, Portugal.

Brewster, C., J. Iria, F. Ciravegna, and Y. Wilks (2005). The Ontology: Chimaera or Pegasus. In *Proceedings of the Dagstuhl Seminar Machine Learning for the Semantic Web*, Dagstuhl, Germany.

Brin, S. (1998). Extracting patterns and relations from the World Wide Web. In *WebDB Workshop at the 6th International Conference on Extending Database Technology (EDBT'98)*.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pp. 120–126.

Chinchor, N. A. (1998). Overview of MUC-7/MET-2. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Choi, K.-S. and H.-S. Bae (2004). Procedures and problems in Korean-Chinese-Japanese WordNet with shared semantic hierarchy. In *Global WordNet Conference*, Brno, Czech Republic.

Cimiano, P., A. Hotho, and S. Staab (2004). Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pp. 435–443.

Cimiano, P. and S. Staab (2004). Learning by googling. *SIGKDD Explorations 6*(2), 24–34.

Cimiano, P. and S. Staab (2005). Learning concept hierarchies from text with a guided agglomerative clustering algorithm. In *Proceedings of the ICML 2005 Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods (OntoML 05)*, Bonn, Germany.

Dagan, I., F. C. N. Pereira, and L. Lee (1994). Similarity-based estimation of word co-occurrence probabilities. In *Meeting of the Association for Computational Linguistics*, pp. 272–278.

de Saussure, F. (1916). *Cours de linguistique générale*. Paris: Payot.

Deerwester, S. C., S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science 41*(6), 391–407.

Ding, Y. and S. Foo (2002). Ontology research and development: Part 1 – A review of ontology generation. *Journal of Information Science 28*(2), 123–136.

Dornseiff, F. (2004). *Der deutsche Wortschatz nach Sachgruppen. 8., völlig neu bearb. u. mit einem vollständigen alphabetischen Zugriffsregister versehene Aufl. von Uwe Quasthoff*. Berlin, New York: Walter de Gruyter.

Eco, U. (1977). *A Theory of Semiotics*. London: The Macmillan Press.

Faure, D. and C. Nédellec (1998). ASIUM: Learning subcategorization frames and restrictions of selection. In Y. Kodratoff (Ed.), *Proceedings of 10th Conference on Machine Learning (ECML 98): Workshop on Text Mining*, Chemnitz, Germany.

Fleischman, M. and E. Hovy (2002). Fine grained classification of named entities. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan.

Gale, W. A., K. W. Church, and D. Yarowsky (1993). A method for disambiguating word senses in a large corpus. *Computers and the Humanities 26*, 415–439.

Gliozzo, A., C. Giuliano, and C. Strapparava (2005). Domain kernels for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, Michigan, USA, pp. 403–410.

Gómez-Pérez, A. and D. Manzano-Macho (2003). A survey of ontology learning methods and techniques. Deliverable 1.5, OntoWeb Project.

Grishman, R. (1997). Information extraction: Techniques and challenges. In *SCIE*, pp. 10–27.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition 5*(2), 199–220.

Harris, Z. S. (1968). *Mathematical Structures of Language*. New York: Interscience Publishers John Wiley & Sons.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING 1992)*, Volume 2, Nantes, France, pp. 539–545.

Heinrich, G., J. Kindermann, C. Lauth, G. Paaß, and J. Sanchez-Monzon (2005). Investigating word correlation at different scopes – a latent topic approach. In *Proceedings of the Workshop on Learning and Extending Lexical Ontologies by using Machine Learning (OntoML 05)*, Bonn, Germany.

Helbig, H. (2001). *Die semantische Struktur natürlicher Sprache*. Heidelberg: Springer.

Heyer, G., U. Quasthoff, and T. Wittig (2005). *Wissensrohstoff Text*. Bochum: W3L-Verlag.

Heyer, G. and H. F. Witschel (2005). Terminology and metadata – on how to efficiently build an ontology. *TermNet News – Newsletter of International Cooperation in Terminology 87*.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Meeting of the Association for Computational Linguistics*, pp. 268–275.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of Uncertainty in Artificial Intelligence*, Stockholm, Sweden, pp. 289–296.

Hoste, V. (2005). *Optimization Issues in Machine Learning of Coreference Resolution*. Ph. D. thesis, University of Antwerp, Belgium.

Husserl, E. (1975). *Logische Untersuchungen 1: Prolegomena zur reinen Logik*. Husserliana 18 (edited by E. Holenstein). Den Haag.

Johansson, C., A. Nøklestad, and C. Biemann (2005). Why the monkey ate the banana. In *Proceedings of the Workshop on Anaphora Resolution (WAR)*, Mjølfjell, Norway.

Keller, F., M. Lapata, and O. Ourioupina (2002). Using the web to overcome data sparseness. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, USA, pp. 230–237.

Kilgarriff, A. (1998). SENSEVAL: An exercise in evaluating word sense disambiguation programs. In *Proceedings of LREC 1998*, Granada, Spain, pp. 581–588.

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

Leopold, E. (2005). On semantic spaces. *LDV-Forum (Special Issue on Text Mining) 20*(1), 63–86.

Leveling, J. and S. Hartrumpf (2005). University of Hagen at CLEF 2004: Indexing and translating concepts for the GIRT task. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini (Eds.), *CLEF 2005*, pp. 271–282. Berlin: Springer.

Maedche, A. and S. Staab (2002). Measuring similarity between ontologies. In *Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW-2002)*, Berlin, pp. 251–263. Springer.

Maedche, A. and S. Staab (2004). Ontology learning. In S. Staab (Ed.), *Handbook on Ontologies*, pp. 173–190. Springer.

Miller, G. A. (1990). WordNet – an on-line lexical database. *International Journal of Lexicography 3*(4), 235–244.

Omelayenko, B. (2001). Learning of ontologies for the web: the analysis of existent approaches. In *Proceedings of the International Workshop on Web Dynamics*.

Paaß, G., J. Kindermann, and E. Leopold (2004). Learning prototype ontologies by hierarchical latent semantic analysis. In *Knowledge Discovery and Ontologies (KDO-2004)*, Pisa, Italy.

Paşca, M. (2005). Finding instance names and alternative glosses on the Web: WordNet reloaded. In *Proceedings of Computational Linguistics and Intelligent Text Processing: 6th International Conference (CICLing 2005)*, LNCS 3406, Mexico City, Mexico, 2005, pp. 280–292.

Pantel, P., D. Ravichandran, and E. Hovy (2004). Towards terascale knowledge acquisition. In *Proceedings of the 20th Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland.

Pease, A. and I. Niles (2002). IEEE standard upper ontology: a progress report. *Knowledge Engineering Review, Special Issue on Ontologies and Agents 17*(1), 65–70.

Pereira, F., N. Tishby, and L. Lee (1993). Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 183–190.

Quine, W. V. (1969). *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Rapp, R. (2002). The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan.

Reuters Corpus (2000). Volume 1, English language, 1996-08-20 to 1997-08-19, release date 2000-11-03, format version 1. `http://about.reuters.com/researchstandards/corpus`.

Rieger, B. B. (1981). Feasible fuzzy semantics. On some problems of how to handle word meaning empirically. In H. Eikmeyer and H. Rieser (Eds.), *Words, Worlds, and Contexts. New Approaches in Word Semantics (Research in Text Theory 6)*, pp. 193–209. Berlin/New York: de Gruyter.

Riloff, E. and R. Jones (1999). Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of AAAI-99*, pp. 474–479.

Riloff, E. and J. Shepherd (1997). A corpus-based approach for building semantic lexicons. In C. Cardie and R. Weischedel (Eds.), *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP 1997)*, Somerset, New Jersey, USA, pp. 117–124. Association for Computational Linguistics.

Ritter, J. and K. Gründer (Eds.) (1995). *Historisches Wörterbuch der Philosophie*. Basel/Stuttgart: Schwabe.

Roark, B. and E. Charniak (1998). Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In *Proceedings of the 36th Annual Meeting of the ACL*, pp. 1110–1116.

Roget, P. (1852). Roget's thesaurus of english words and phrases. In *Longman*, London.

Ruge, G. (1992). Experiment on linguistically-based term associations. *Information Processing and Management 28*(3), 317–332.

Salton, G., A. Wong, and C. S. Yang (1975). A vector space model for automatic indexing. *Communications of the ACM 18*(11), 613–620.

Sánchez, D. and A. Moreno (2005). Web-scale taxonomy learning. In *Proceedings of the Workshop on Learning and Extending Lexical Ontologies by using Machine Learning (OntoML 05)*, Bonn, Germany.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics 24*(1), 97–123.

Smith, B. (2004). Ontology. In L. Floridi (Ed.), *The Blackwell Guide to Philosophy of Computing and Information*. Blackwell: Malden.

Sowa, J. F. (2003). Ontology. `http://www.jfsowa.com/ontology/` (last changed 2003).

Staab, S., C. Braun, A. Düsterhöft, A. Heuer, M. Klettke, S. Melzig, G. Neumann, B. Prager, J. Pretzel, H.-P. Schnurr, R. Studer, H. Uszkoreit, and B. Wrenger (1999). GETESS – Searching the web exploiting german texts. In *CIA'99: Proceedings of the Third International Workshop on Cooperative Information Agents III*, London, UK, pp. 113–124. Springer.

Stamou, S. and D. Christodoulakis (2005). Retrieval efficiency of normalized query expansion. In *Proceedings of Computational Linguistics and Intelligent Text Processing: 6th International Conference (CICLing 2005)*, LNCS 3406, Mexico City, Mexico, pp. 593–596.

Thelen, M. and E. Riloff (2002). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, USA.

Vossen, P. (1997). EuroWordNet: A multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997*, Zürich, Switzerland.

Wagner, A. (2000). Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis. In *Proceedings of the ECAI 2000 Workshop on Ontology Learning*, Berlin, Germany.

Widdows, D. (2003). Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *HLT-NAACL 2003: Main Proceedings*, pp. 276–283.

Witschel, H. F. (2005). Using decision trees and text mining techniques for extending taxonomies. In *Proceedings of the Workshop on Learning and Extending Lexical Ontologies by using Machine Learning (OntoML 05)*, Bonn, Germany.