The Status Quo of Ontology Learning from Unstructured Knowledge Sources for Knowledge Management

Andreas Scheuermann[1] and Jens Obermann[2]

[1] University of Hohenheim, Information Systems 2, Schwerzstraße 35,
70599 Stuttgart, Germany
andreas.scheuermann@uni-hohenheim.de
[2] SAP Deutschland GmbH & Co. KG, Hasso-Plattner-Ring 7
69190 Walldorf, Germany
jens.obermann@sap.com

Abstract. In the global race for competitive advantage Knowledge Management gains increasing importance for companies. The purposeful and systematic creation, maintenance, and transfer of unstructured knowledge sources demands for advanced Information Technology. Ontologies constitute a basic ingredient of Knowledge Management; thus, ontology learning from unstructured knowledge sources is of particular interest since it bears the potential to bring significant advantages for Knowledge Management. This paper presents a study of state-of-the-art research of ontology learning from unstructured knowledge sources for Knowledge Management. Nine approaches for ontology learning from unstructured knowledge sources are identified from a systematic review of literature. A six point classification framework is developed. The review results are analyzed, synthesized, and discussed to give an account of the current state-of-the-art for contributing to an enhanced understanding of ontology learning from unstructured knowledge sources for Knowledge Management.

Keywords: Ontology Learning, Knowledge Management, Literature Review.

1 Introduction

In the global race for competitive advantage the ultimate success of companies increasingly depends on effectively and efficiently exploiting knowledge. This knowledge is often distributed, heterogeneous, and contained in various types of knowledge sources. The purposeful and systematic creation, maintenance, and transfer of knowledge are subject of Knowledge Management (KM). KM depends on non-technical aspects but increasingly draws upon the use of technical properties such as advanced Information Technology (IT). Such IT could significantly benefit from the use of Semantic Technology and particularly ontologies provide dedicated means to enhance the future challenges of Knowledge Management [1-2].

However, constructing ontologies (from scratch) is a non-trivial and complex task, which requires considerable efforts regarding costs, time, and labor. Ontology construction involves experts from the area of ontology engineering and experts from the particular domain of interest. Experts are typically scarce and acquiring the relevant

E. Mercier-Laurent and D. Boulanger (Eds.): AI4KM 2012, IFIP AICT 422, pp. 72–94, 2014.

knowledge from a human domain expert is rather difficult due to the nature of human knowledge (e.g., implicit and procedural knowledge) and miscommunications. Despite ontologies bear the potential to significantly contribute to the further advancement of Knowledge Management, there are several obstacles specifically with regard to ontology construction, extension, and refinement that constrain a widespread adoption and use of ontologies in KM.

To overcome these obstacles, ontology learning represents a promising but yet not fully exploited approach to enable the (semi-)automatic construction, extension, and refinement of ontology. Ontology learning principally allows to (semi-) automate parts of the non-trivial and complex task of ontology construction and, thus, to significantly reduce cost, time, and labor expenses. That is, ontology learning appears to be an ideal solution to leverage ontologies for advancing KM not only from a technical but also from an economic perspective.

Unfortunately, huge amounts of knowledge exhibit a high relevance for companies but their representational form lacks a clear structure and organization [3]. This unstructured knowledge aggravates (semi-)automatic machine-readability and interpretability; thus, the use of approaches such as ontology learning. For instance, the wide adoption of Social Media on the World Wide Web (WWW) increased the number of unstructured knowledge sources dramatically. Social Media in terms of Facebook, Twitter, Blogs, and further types of these applications contain plenty of unstructured knowledge, which can be of major significance for company purposes such as marketing, product development, consumer studies, customer relationships, advertising, recruiting, etc. Despite the inherent characteristic features of this knowledge type hamper the use of (semi-)automatic approaches, increase the number of errors and demand for human intervention, exploiting semi- and unstructured knowledge sources for reasons of competitive advantage gains more and more significance.

The objective of this paper is to give an account of the current state-of-the-art in order to contribute to an enhanced understanding of ontology learning from unstructured knowledge sources for Knowledge Management. Therefore, this paper describes, analyzes, and assesses extant ontology learning approaches with regard to their capabilities to process unstructured knowledge sources for reasons of Knowledge Management. Thus, this paper extends and refines [4-6] in the matter of unstructured knowledge sources and Knowledge Management.

The remainder of this paper is organized as follows: Section 2 introduces ontology learning for knowledge management, which covers basic constructs from the areas of Knowledge Management, ontology, and ontology learning. Section 3 characterizes the review strategy, presents the classification framework, and introduces the identified ontology learning approaches. Section 4 presents the review results and critically reflects on them. Finally, Section 5 draws the conclusion.

2 Ontology Learning for Knowledge Management

2.1 Knowledge Management

Knowledge Management gains increasing importance for the competitiveness of companies and, thus attracted growing attention from both industry and academia.

KM essentially deals with cultural, organizational, human, and technical issues covering three processes of creating, maintaining, and transferring various forms of knowledge in an intra- and inter-organizational context [7]. These three overarching processes can be further decomposed into a coherent and consistent set of six Knowledge Management core processes [8]: (1) knowledge identification, (2) knowledge acquirement, (3) knowledge development, (4) knowledge distribution, (5) knowledge use, and (6) knowledge preservation (Fig. 1).

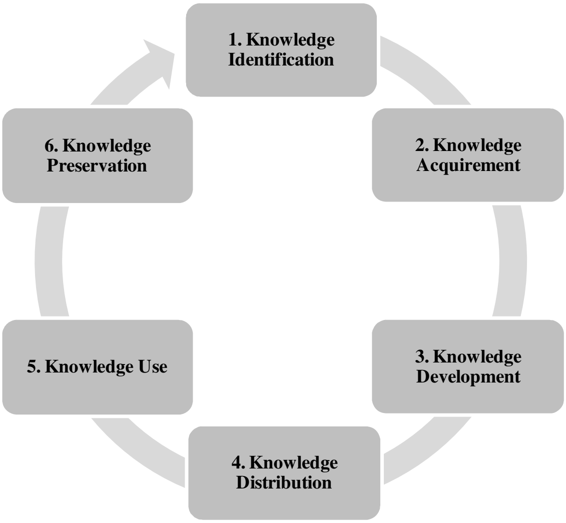
1. Knowledge Identification
2. Knowledge Acquirement
3. Knowledge Development
4. Knowledge Distribution
5. Knowledge Use
6. Knowledge Preservation

Fig. 1. Knowledge Management Core Processes

Subsequently, the constructs constituting the Knowledge Management core processes are briefly characterized:

1. *Knowledge identification* concerns the characterization and specification of the knowledge relevant for the organization.
2. *Knowledge acquirement* elicits and collects the identified knowledge.
3. *Knowledge development* augments and combines the acquired knowledge for creating new knowledge.
4. *Knowledge distribution* provides appropriate means to distribute and make the knowledge available.
5. *Knowledge use* depicts the exploitation of the knowledge for accomplishing the organizational goals.
6. *Knowledge preservation* is responsible for long-term knowledge storage.

Despite various cultural, organizational, and human factors exert an influence on the Knowledge Management core processes, the focus is on technical issues of KM and specifically ontology learning. To elaborate on ontology learning for KM, it is reasonable to understand the potential roles and benefits of ontologies in KM.

2.2 Ontology

The term ontology originates from philosophy and denotes the discipline of philosophy that concerns “the science of what is, of the kinds and structures of objects, properties events, processes, and relations in every area of reality” [9]. With the beginning of the late 1980s and early 1990s, ontology attracted growing attention and became subject of research in Computer Science and Artificial Intelligence (AI). AI research deals with formal representations of real world phenomena and reasoning about these phenomena. In a literal sense, AI research *borrowed* the term ontology from philosophy [10], equipped it with a computational meaning, and, thus, coined the term “formal ontology” (or computational ontology).

In the following, AI research studied ontology for its purposes, most notably in the context of knowledge engineering and knowledge representation, and contributed to several definitions. The most prominent definition stems from [10-11] and defines ontology as “an explicit specification of a conceptualization”. A refinement of this definition provides [11] in terms of requiring the specification to be formal and the conceptualization to be shared. Based on this, [13] concisely and comprehensively points out the characteristics features of formal ontology by defining ontology as “a formal, explicit specification of a shared conceptualization of a domain of interest”.

- C*onceptualization* depicts an abstract representation of some (real-world) phenomenon by having determined its relevant concepts, relations, axioms, and constraints.
- E*xplicit* denotes the explicit (not implicit) definition of the type of concepts and relations as well as the axioms and constraints holding on their use.
- *Formal* indicates that the ontology should be readable and interpretable by machines, thus, formal excludes the use of natural language.
- S*hared* conceptualization reflects that an ontology captures consensual knowledge that is not private to an individual person but accepted by a larger group of individuals.

For reasons of the formal and explicit representation of consensual knowledge about a particular domain of interest, ontology draws upon the following set of basic modeling primitives [11]:

- *Classes* typically follow a hierarchical organization, which allows for applying inheritance mechanisms. Classes are used in a broad sense (types of anything); thus, they can be either abstract (e.g., intentions or beliefs), concrete (e.g., people or trees), elementary, or composite.
- *Relations* define the type of associations between classes and essentially distinguish between three types of relations: (1) unary relations, (2) binary relations, and (3) functions.
- *Axioms* model true statements. Ontology contains axioms (1) to constrain the knowledge, (2) to verify the correctness of the knowledge, and (3) to deduce new knowledge.

- *Instances* represent elements of a specific class whereas facts depict the relation between elements. Both instances and facts, i.e. any element of a domain of interest that is not a class refers to as individuals.

Against this background, the potential applications of ontologies spans a wide array but primarily aim at knowledge sharing and reuse, which includes (1) the formal specification of knowledge, (2) the structuring and organization of knowledge, and (3) the provision of a common terminology, i.e. interlingua [10],[13-16]. Regarding KM, the potential uses and roles of ontology primarily concern the Knowledge Management core process of:

- knowledge acquirement,
- knowledge development,
- knowledge distribution, and
- knowledge use.

Based on the definition of ontology and the clarification of its potential roles and benefits within KM, the subsequent section deals with ontology learning to narrow the scope and carve out its potential use and benefits for KM.

2.3 Ontology Learning

Ontology engineering is defined “as the set of activities that concern the ontology development process, the ontology life cycle, and the methods, tools, and languages for building ontologies” [17-18]. A closer inspection of this definition indicates that ontology engineering essentially covers the following five areas:

- the ontology development process,
- the ontology lifecycle,
- the methods for developing ontologies,
- the tools that support ontology development, and
- the (ontology) languages, which are applied in ontology development.

For elaborating on ontology learning, it is reasonable to focus on both the ontology development process and methods for developing ontologies. In particular, ontology development distinguishes between several different types of methods that deal with the creation of ontologies based on specific approaches: (1) methods for developing ontology from scratch, (2) methods for re-engineering of existing ontologies, (3) methods for ontology alignment and ontology merging, and (4) methods for ontology learning [17-18].

- Ontology development from scratch deals with newly developing ontologies, which means that large parts of the envisioned ontology are manually constructed whereas ontology reuse plays a minor role. Developing ontologies from scratch is required when ontologies of the particular domain of interest lack quality, availability, coverage, or not yet exist.

- Ontology re-engineering adapts preexisting ontologies according to specific requirements and covers the following steps: (1) retrieve the conceptualization of an ontology implementation, (2) transform it, i.e. extend, refine, and prune this conceptualization according to the given requirements, and (3) re-implement the (re-engineered) ontology.
- Ontology alignment (matching) and ontology merging methods generally aim at unifying preexisting ontologies. In particular, ontology alignment establishes various kinds of mappings between the ontologies and, thus, preserves the original ontologies. In contrast to that, ontology merging generates a unified ontology from the original ontologies but does not preserve the original ontologies.
- Ontology learning (semi-)automatically acquires knowledge from knowledge sources to create, enrich, or populate ontologies. Ontology learning typically draws upon preexisting knowledge structures such as taxonomies or ontologies that already capture parts of the domain of interest and presuppose the existence and availability of external knowledge sources.

This classification depicts that ontology learning bears the potential to significantly contribute to the further advancement of Knowledge Management. Ontology learning is capable to overcome impediments related to the construction, extension, and refinement of ontologies with regard to costs, time, and human labor [4-5]. In particular, ontology learning has the potential to advance the following knowledge management core processes:

- knowledge acquirement and
- knowledge development.

However, to better understand the use and potentials of ontology learning for the two knowledge management core processes, it is reasonable to characterize ontology learning in more detail and classify existing approaches for ontology learning.

Characteristics

Ontology learning generally aims at automating ontology development with regard to acquiring knowledge from several types of knowledge sources and, then, constructing an ontology or components of an ontology (e.g., classes or properties). Therefore, ontology learning primarily draws upon constructs, methods, and techniques from the field of information extraction to elicit information, patterns, and relations from various kinds of knowledge sources as well as ontology engineering to construct the ontology or its components. In essence, ontology learning pursues reductions in cost, time, and human labor, e.g., in terms of a reduced level of human interaction. [20],[25].

From a technical point of view, ontology learning combines the following two constituent components: (1) information extraction approaches and (2) (ontology) learning approaches.

First, ontology learning principally distinguishes between two distinct approaches for information extraction: (1) rule-based approaches and (2) heuristics pattern approaches [25]. Both approaches draw upon lexico-syntactic patterns, which allow for dealing with knowledge sources characterized by insufficient pre-encoded knowledge and wide ranges of data (e.g., text) [21]. Lexico-syntactic patterns aim at finding recurring and recognizable structures within data sets (e.g., text) without depending on sets of fixed terms of expressions, which need to be determined a priori.

Second, ontology learning incorporates learning algorithms, which essentially assess whether the extracted information entities fit the purpose and scope of the envisioned ontology. In case of a positive assessment, the learning algorithm adds the information entity as a new element to the ontology, that is, in terms of an ontology component [6],[22].

In principle, ontology learning differentiates from other existing approaches by its multidisciplinary applicability and the potential to exploit vast and heterogeneous knowledge sources [23].

Process

Ontology learning principally combines and integrates information extraction and ontology learning approaches. In addition to that, ontology learning approaches exploit the outcomes of numerous disciplines such as linguistics, statistics, heuristic and pattern matching, machine learning, or data mining for applying them on various types of knowledge sources to enhance the quality of the results. Based on that, ontology learning is a complex and multifaceted process. To point out the characteristics features of the ontology learning process, this paper refers to the ontology learning process proposed by [24].

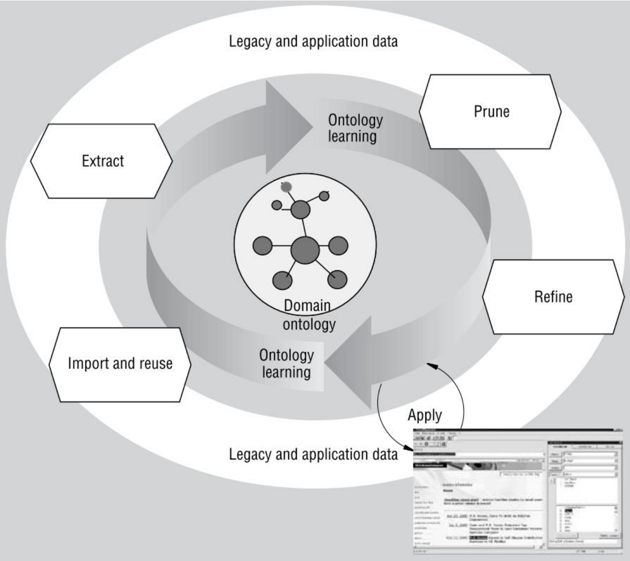

Fig. 2. Ontology Learning Process [24]

The ontology learning process shown in Fig. 3 is rather generic and encompasses the following four main activities:

- *Import and reuse* deals with mapping and merging preexisting knowledge structures.
- *Extract* concerns the creation of major parts of the ontology based on information extraction from various types of knowledge sources.
- *Prune* aims at tailoring the preliminary ontology to satisfy its purpose and scope according to specific requirements of the target application.
- *Refine* considers the completion of the ontology at a more detailed and fine-grained level.

After performing these four main activities of the ontology learning process, the target application provides a testbed and measure for validating and further refining the envisioned ontology. Further, this ontology learning process allows the ontology engineer to extend the ontology, i.e. to perform further iterations of the ontology learning process as well as to update and maintain the ontology in the course of the ontology lifecycle.

Classification

Ontology learning is an active area of research and, in accordance to that, literature yields several schemes for classifying ontology learning approaches based on multiple criteria such as type of preprocessing, preexisting knowledge structures, learning algorithms, learned ontology components, or degree of human interaction. In accordance to the purpose and scope of this paper, it is reasonable to classify ontology learning with regard to different types of knowledge sources. For instance, [24] classify ontology learning according to the following types of knowledge sources: free texts, dictionaries, semi-structured schemata, and relational schemata. Influenced by [24], this work categorizes ontology learning approaches based on the following three types of knowledge sources: (1) structured knowledge sources, (2) semi-structured knowledge sources, and (3) unstructured knowledge sources.

Next, each of these types of knowledge sources is briefly characterized:

- *structured knowledge sources* are tightly coupled to specific rules of a conceptualization, e.g., relational databases and the relational schema.
- *semi-structured knowledge sources* incorporate some rules of a conceptualization but also contain unstructured elements, e.g., HTML documents
- *unstructured knowledge sources* can be of any kind and lack particular rules or structure. The key characteristic features of unstructured knowledge sources are the high availability throughout all domains but also the lowest accessibility for ontology learning.

Against the background of these three types of knowledge sources, the scope of this paper covers semi-automatic and automatic ontology learning approaches from unstructured knowledge sources.

3 Methodology

3.1 Review Strategy

To identify the relevant ontology learning approaches in the literature, a structured and iterative approach was employed. A systematic search was used to retrieve publications (journal articles, conference and workshop proceedings, as well as technical reports) that reported on ontology learning. The search combines both a keyword-based and explorative search strategy to reach a maximum coverage and to accomplish high quality search results. The keyword-based search assembled multiple search strings (e.g., ontology learning, unstructured knowledge, etc.) in various forms for searching online databases (e.g., Google Scholar, Scopus, ACM Digital Library, etc.). The search strategy relied on citation count as a proxy measure to identify probable core publications. Since filtering based on citation count may exclude some relevant ontology learning approaches, an explorative search was used to find additional publications on ontology learning. Furthermore, the search strategy was iterative to both reduce the list of search results (e.g., adding constraints to the search query) as well as expand the list (e.g., adding alternative terms to the search query). Employing this review strategy resulted in a manageable list of potentially relevant ontology learning approaches. This list was then manually inspected by analyzing the abstracts and skimming the content, resulting in nine publications.

3.2 Classification Framework

The classification framework consists of six criteria for describing and analyzing the indentified ontology learning approaches. These six criteria stem from an analysis of literature in the area of ontology learning and are assessed relevant for reviewing ontology learning approaches with respect to unstructured knowledge sources and Knowledge Management. As such, these criteria are extrinsic in their nature and allow for an assessment from an objective point of view. In particular, the various criteria reflect descriptive constructs of the domain of ontology learning that focus both on methodological properties of the ontology learning approaches (Criteria 1-5) as well as on the resulting ontologies (Criterion 6). In the following, Table 1 depicts the classification framework and the six constituent criteria with example concrete measurements before each of these criteria is briefly characterized.

Table 1. Classification Framework

#	Criterion	Concrete Measurements
1	Objective	Purpose, scope, target application
2	Methodology	Statistical, logical, natural language
3	Technique	Supervised, unsupervised
4	Degree of Automation	Semi-automatic, (fully) automatic
5	Reuse of Knowledge Sources	Lexica, taxonomies, ontology
6	Ontology Components	Classes, relations, instances, taxonomies

Criterion 1: Objective
(Objective) aims to detect and analyze the primary goal of ontology learning in terms of the purpose and scope of the envisioned ontology based on the specification of the identified problem in the target application.

Criterion 2: Methodology
(Methodology) aims to detect and analyze the methodological approach that underpins ontology learning from unstructured knowledge. This criterion pays special attention to statistical and natural language processing (NLP) approaches for ontology learning as they frequently occur in literature and can be assessed promising for Knowledge Management.

Criterion 3: Technique
(Technique) aims to detect and analyze the technique for extracting information and learning; thus, this criterion specializes Criterion 2. For instance, it elaborates on the process of ontology learning in terms unsupervised or supervised.

Criterion 4: Degree of Automation
(Degree of Automation) aims to detect the degree of automation at which the ontology learning approach is supposed to operate. The degree of automation basically distinguishes between semi-automatic and (fully) automatic approaches for ontology learning and relates to economical advantages, which gain importance for knowledge management in fast changing business environments.

Criterion 5: Reuse of Knowledge Sources
(Reuse of knowledge Sources) aims to detect whether the ontology learning approach reuses preexisting (formal) bodies of knowledge, e.g., WordNet. This is especially interesting for KM as it can be assumed that there are already ontologies, which have to be extended or refined according to changing business demands. In addition to that, this criterion expresses the basic understanding of the ontology learning approach: building up ontologies from scratch or finding a preexisting knowledge structure as the starting point.

Criterion 6: Ontology Components
(Ontology components) aim to detect and analyze the ontology components, which essentially correspond to the results from information extraction and, then, become the subjects of learning. Ontology components typically correspond to classes, relations, axioms, and instances but might also cover taxonomies or further types of knowledge structures. This criterion highlights the envisioned results of ontology learning and maintains close relationships to Criterion 3 and Criterion 5. Criterion 6 is of particular importance for KM since it provides information about the envisioned ontology.

3.3 Identified Ontology Learning Approaches

Ontology learning from unstructured knowledge sources primarily distinguishes between two different types of approaches: (1) statistical approaches and (2) natural language processing (NLP) approaches. The search result finally comprises four statistical ontology learning approaches and five ontology learning approaches based on NLP (Table 2).

Table 2. Identified Ontology Learning Approaches

Authors	Type	Short Description
Agirre et. al (2000) [25]	Statistical	Enriching very large ontologies using the WWW
Faatz and Steinmetz (2002) [26]	Statistical	Ontology enrichment with texts from the WWW
Sanchez and Moreno (2004) [27]	Statistical	Creation of ontologies from web documents
Cimiano et al. (2005) [28]	Statistical	Learning of concept hierarchies from text corpora using formal concept analysis
Hearst (1992) [21]	NLP	Automatic acquisition of hyponyms from large text corpora
Kietz et al. (2000) [29]	NLP	Semi-automatic ontology acquisition from corporate intranets.
Gupta et al. (2002) [30]	NLP	Architecture for engineering sublanguages – WordNets
Alfonseca and Manandhar (2002) [31]	NLP	Extension of a lexical ontology by a combination of distributional semantics signatures
Narr et al. (2011) [32]	NLP	Extraction of semantic annotations from Twitter

In the following, the review describes and analyzes the nine ontology learning approaches based on the classification in statistical and NLP approaches as well as in accordance to the chronological order of the year of development.

Statistical Approaches

Statistical ontology learning approaches that deal with unstructured knowledge sources draw upon a common basic assumption. This basic assumption corresponds to the distributional hypothesis [33]. The distributional hypothesis states that similar words often occur in similar contexts and, thus, statistical patterns provide hints for certain relations between words.

Agirre et al. (2000) Approach

[25] propose an automatic ontology learning approach, which deals with enriching the WordNet database by using unstructured knowledge sources, i.e. the WWW. The enrichment of WordNet is deemed necessary because of two major drawbacks: (1) semantic variant concepts of words, which are related by topics, are not interlinked (e.g., to paint and paint or sun cream and beach) and (2) the vast collection of word meanings without any clear distinction. In this context, the proposed ontology learning approach primarily relies on word lists. Word lists describe the sense of the words

of interest. This sense is based on the idea that other specific words describe the context and, thus, express the meaning of the word of interest (indirectly and implicitly). Initially, the proposed approach queries (boolean search query) the WWW for documents (datasets), which contain the word of interest. Thereby, the higher the number of the words of interest in the retrieved documents, the higher the statistical likelihood that the document correlates with the topic searched for. To increase this likelihood, the search could explicitly include and exclude further descriptive constructs, i.e. words. Then, counting the occurrence of single words in the documents and using calculated distance metrics to hierarchically sort them results in topic signatures. These topic signatures are clustered and evaluated by means of a disambiguation algorithm. Thereby, clustering corresponds to a common technique to generate prototype-based and hierarchical ontologies. A predefined semantic distance algorithm works as a measurement to agglomerate terms or clusters of terms. The largest or the least homogeneous cluster is split into smaller subgroups by a divisive process to further refine the envisioned ontology.

Faatz and Steinmetz (2002) Approach

[26] introduce a semi-automatic ontology learning approach, which primarily aims at enriching preexisting ontologies. The proposed approach is illustrated with an example of a medical ontology. Similar to [25], [26] use a statistical approach to cluster words, which occur in a certain context to each other. In addition to that, this approach incorporates a set of predefined rules. This set of rules represents distance measurements, e.g., maximum word distance between two words in a document, which, in principle, should not be contradictory to already existing distance measures. The statistical similarity measurement generally relies on the Kullback-Leibler divergence [34], which was originally designed to test the probability distribution between statistical populations in terms of information. In particular, [26] use the Kullback-Leibler divergence to check the weighted probability of a given linguistic property w with respect to its fulfillment by a word x. This allows for assessing and minimizing the distance of the word of interest and the retrieved document in a way similar to an optimization problem.

Sanchez and Moreno (2004) Approach

[27] develop an automatic ontology learning approach, which exploits web documents as the primary knowledge source to create ontologies. Using the WWW for ontology creation might increase the probability that the ontology reflects the current state of practice or knowledge and, thus, is more complete. Similar to [25] and [26], [27] formulate queries to search for specific words of interest and constraint this search by criteria like the maximum number of returned results as well as by the use of filters for similar documents. Based on the initial search results, a first analysis according to predefined prerequisites is conducted for filtering relevant documents. Then, the application of statistical analysis techniques aims at further filtering the most relevant documents from this subset. The next step is to filter the results from the previous step by adding a new search word to refine the original search. The goal of the last two

steps is mainly to increase search depth. Moreover, the resulting taxonomies support finding new relations between ontologies.

Cimiano et al. (2005) Approach
[28] propose the adoption of Formal Concept Analysis (FCA) for ontology learning on an automated basis. The proposed ontology learning approach analysis documents by searching for sets of object attributes and, based on them, derives relationships and dependencies between the objects. The results of this search conform to nouns associated with several verbs as trailed attributes. These attributes define the context of the noun. The formal abstraction of the inherited nouns provides additional benefits to an end-user as the verbs provide a further foundation to enrich the envisioned ontology. The reason for this is potentially because of the more adequate description by a verb in contrast to a noun or hyponym.

NLP Approaches
Prior to characterizing the identified ontology learning approaches based on NLP, it is reasonable to note that literature lacks a clear and consensual distinction between statistical and NLP approaches for ontology learning. Despite ontology learning approaches draw upon statistics and linguistics to exploit unstructured knowledge sources, NLP approaches incorporate a more intuitive way of dealing with unstructured knowledge sources by using techniques such as pattern recognition [35]. In particular, ontology learning approaches based on NLP provide additional benefits with regard to knowledge-intensive domains, which require several constraints and rules within ontology learning. Such constraints and rules conform to lexical inventories, syntactic rules, or predefined knowledge structures [36].

Hearst (1992) Approach
[21] introduces an ontology learning approach, which uses the lexico-syntactic pattern extraction method to support the enrichment of preexisting patterns within the WordNet database by searching large text corpora as a mining resource for suitable semantic patterns. A crucial prerequisite of this approach is that the English language has identifiable lexico-syntactic patterns, which indicate specific semantic relations in terms of is-a relations. In comparison to other approaches, the underlying text corpora have to fulfill only very little usage requirements. In particular, this means that only one instance of a relation has to be available in the knowledge source to decide whether the knowledge source is suitable or not. [21] uses a deterministic system to provide one or several hyponyms for each unknown concept, which all have a certain probability to be correct based on the unstructured knowledge source. To increase the suitability of the derived concepts, the lexico-syntactic patterns have to fulfill some criteria. That is, the lexico-syntactic patterns (1) need to frequently occur in the text corpora, (2) indicate the relation of interest, and (3) allow for a potential recognition without any prior knowledge of the domain of interest. Moreover, this approach allows for combinations with other techniques such as statistical algorithms for the purpose of refining the patterns.

Kietz et al. (2000) Approach

[29] essentially build on the approach proposed by [21] to introduce an ontology learning approach that allows for semi-automatically create ontologies from text corpora retrieved from corporate intranets. The proposed approach incorporates a learning method, which is based on a set of given core concepts similar to WordNet. This learning method further uses statistical and pattern-based techniques to refine the respective results. Then, the resultant ontology is pruned and restricted. In comparison to prior approaches, this (non-taxonomic) approach uses conceptual relations rather than manually encoded rules for the purpose of ontology creation. Moreover, this approach comprises a set of evaluation metrics to ensure a hegemonic ontology with respect to the target knowledge structure. However, these metrics are not conclusive enough to fully automate the process of ontology learning.

Gupta et al. (2002) Approach

[30] introduce an ontology learning approach, which primarily aims at speeding up the ontology learning process by using WordNet and, particularly, by creating sublanguages, i.e. so-called WordNets. The creation of these sublanguages results from an application of acronym extractors and phrase generators for analyzing knowledge sources for concept elements. For instance, concept elements are words, potential relations between words, and phrases. Then, potential relationships are analyzed again and proposed as candidates for being added to the envisioned ontology. Thereby, words and suitable relationships are clustered into groups and linked to the corresponding synsets in WordNet as WordNets. Finally, the last step focuses on maintenance of the retrieved concept elements and knowledge structures.

Alfonseca and Manandhar (2002) Approach

[31] introduce an automatic ontology learning approach to extend a lexical semantic ontology. The proposed algorithm searches the existing ontology for similarity to a synset for information extraction. For this purpose, [31] define several signatures for the word of interest, which are evaluated with regard to their semantic similarity to existing words within the preexisting ontology. The signatures conform to: (1) topic signatures, which define a list and the frequency of co-occurring words, (2) subject signatures, which inherit a list of co-occurring verbs, (3) object signatures, which contain a list of verbs and prepositions, (4) and modifier signatures, which consist of adjectives and determiners. Thereby, similar words should be assigned with similar signatures since they are represented in similar contexts. All the signatures are aggregated to an overall similarity value. For assessing the frequency, [31] use the method of [25] to achieve more accurate results. Thereto, the plain frequencies are changed into weights, which assess the support of a word in a specific context of a synset.

Narr et al. (2011) Approach

More recent research in the area of NLP extends the area of which text corpora are derived from. In this context, [32] introduce an ontology learning approach to extract information directly from Twitter messages. This approach refines the attempt to retrieve information from unstructured knowledge sources to a new and even more demanding level. Besides the challenge of retrieving the correct semantic relation in the in the unstructured knowledge sources, the problems of misspellings, abbreviations, and colloquial speech in Tweets occurs. Therefore, a normalization of the text is necessary prior to initializing the process of extending or refining the ontology. In addition to enriching an ontology, annotations from the Twitter messages are included in the retrieved semantic structures. These annotations refer to as contextual relations like opinions.

4 Results and Discussion

This section synthesizes, summarizes, and discusses the main results of the review ontology learning approaches from unstructured knowledge for Knowledge Management.

4.1 Results Statistical Approaches

Prior to presenting the review results in detail for each of the four statistical ontology learning approaches, Table 3 provides a summary and overview of the main findings.

Table 3. Results Statistical Ontology Learning Approaches

	Objective; Methodology	**Technique; Degree of Automation**	**Reuse of Knowledge Sources; Ontology Components**
[25]	Ontology enrichment; (primarily) statistics	Harris‘ distributional hypothesis, topic signatures, clustering by statistical measures; automatic	WordNet ontology; classes and relations
[26]	Ontology enrichment; (primarily) statistics	predefined text resources, clustering and statistical information, similarity measures; semi-automatic	medical ontologies; classes and relations
[27]	Ontology enrichment, (primarily) statistics	Keyword-based information by key words; automatic	WWW; Classes
[28]	Ontology creation; (primarily) statistics	Formal Concept Analysis, Harris‘ distributional hypothesis; semi-automatic	domain experts select specific knowledge sources; classes and taxonomies

Agirre et al. (2000)
[25] focus on the problem of word ambiguity in order to enhance ontology learning. To enrich ontologies, [25] use Harris' distributional hypothesis as the foundation for measuring the relevance of the retrieved knowledge elements. These knowledge elements conform to classes and relations. The proposed ontology learning approach can be performed automatically and exploits WordNet as its underlying knowledge source.

Faatz and Steinmetz (2002)
[26] use clustering techniques and similarity measures to perform ontology learning with regard to classes and relations from the retrieved knowledge sources. Instead of using randomly assigned knowledge sources, [26] define several knowledge sources, e.g., documents, which already deal with the topic to provide the basis for the operations of the algorithm. The proposed approach can be categorized as semi-automatic, since suitable knowledge sources are selected a priori by human, i.e. manual intervention. Further, this ontology learning approach mainly reuses medical ontologies with respect to the target application area.

Sanchez and Moreno (2004)
[27] introduce an ontology learning approach that aims at creating new ontologies. Therefore, this approach primarily enriches preexisting ontologies by means of the inclusion of additional knowledge elements, e.g., from the WWW. In general, [27] use the technique of keyword-based information extraction as the first step. Then, the algorithm focuses on the automatic processing of the ontology components, i.e. classes.

Cimiano et al. (2005)
[28] draw upon FCA to create ontologies from scratch. In addition to that, this ontology learning approach builds on Harris' distributional hypothesis. As such, the proposed approach allows for deriving classes and taxonomies from knowledge sources. However, theses knowledge sources have to be manually selected in accordance to the specific topic at hand. As a result, it is reasonable to argue that this ontology learning approach operates semi-automatically.

4.2 Results NLP Approaches

Before presenting the review results in detail for each of the five ontology learning approaches based on NLP, Table 4 provides a summary and overview of the main findings.

Table 4. Results Ontology Learning Approaches based on NLP

	Objective; Methodology	Technique; Degree of Automation	Reuse of Knowledge Sources; Ontology Components
[21]	Ontology enrichment; (primarily) NLP	lexico-syntactic patterns; automatic	WordNet; classes and hyponym-hypernym relations
[29]	Ontology creation; (primarily) NLP	statistical and pattern-based techniques; semi-automatic	GermaNet, WordNet; classes, is-a relations
[30]	Ontology enrichment; (primarily) NLP	Retrieval of sublanguages – WordNets – and adding synsets; semi-automatic	WordNet; WordNets
[31]	Ontology enrichment; (primarily) NLP	Information signatures; automatic	WordNet; objects, classes and hyponym-hypernym relations
[32]	Ontology enrichment; (primarily) NLP	Several NLP techniques; automatic	None; annotations and contextual relations

Hearst (1992)

The ontology learning approach proposed by [21] aims at enriching existing ontologies by exploiting knowledge sources not limited to specific domains. [21] uses lexico-syntactic patterns for retrieving classes and hyponym-hypernym relations. Furthermore, this approach reuses WordNet as its underlying knowledge source. The approach incorporates algorithms that are supposed to work in an automatic way.

Kietz et al. (2000)

[29] introduce an ontology learning approach, which takes an application-driven perspective since its main target corresponds to exploiting companies' intranets as the primary knowledge source. This approach is capable to retrieve classes and *is-a* relations from corporate intranets. Because of the nature of the intranets, this approach operates only semi-automatically. It reuses GermaNet and WordNet as the two major ontologies that underpin ontology learning.

Gupta et al. (2002)

[30] draw upon WordNet but essentially create sublanguages so-called WordNets for the purpose of enriching an upper ontology with regard to more domain-specific knowledge elements. The rationale underpinning this approach is to derive complete WordNets based on the assumption that the domain expert has selected suitable knowledge sources. This approach is capable of updating single WordNets without the need of dealing with the entire ontology. Nevertheless, this approach demands for human intervention and, thus, can only be performed semi-automatically.

Alfonseca and Manandhar (2002)
[31] develop an ontology learning approach, which aims at enriching preexisting ontologies of a domain of interest by exploiting predefined knowledge sources. The proposed approach works automatically without supervision as it generates information signatures based on a set of criteria. In addition to that, carrying out this approach requires large knowledge sources for generating adequate signatures for a unique identification of the envisioned ontology components.

Narr et al. (2011)
[32] propose an ontology learning approach with the goal to enrich preexisting ontologies by exploiting Twitter. The derived ontologies can be enriched with annotations and contextual relationships depending on accompanying words or hash-tags in the Twitter feeds. This approach is supposed to operate completely automatic.

4.3 Discussion

Ontology learning essentially distinguishes between two different approaches, i.e. underpinning methodologies in terms of statistics and NLP. The majority of these approaches aim at ontology enrichment, i.e. extending, refining, or populating preexisting ontologies instead of creating new ontologies from scratch. This observation implies that the majority of the reviewed ontology leaning approaches presupposes an existing knowledge structure, i.e. a taxonomy or an ontology as a starting point. For instance, the ontology learning approaches proposed by [27-28] primarily draw upon statistics and constitute the two exceptional cases. This circumstance may be due to fact that there is a higher likelihood of errors when an ontology is constructed from scratch since there is a severe lack of respective guidelines and possible comparisons to similar structures. In addition to that, [30] introduces an interesting ontology learning approach. That is, [30] aim at the creation of ontologies by means of constructing synsets (sublanguages of WordNet) but these synsets have to be connected to an upper ontology.

Furthermore, there are ontology learning approaches, which aim at the construction of *is-a* relations between classes. However, these approaches rather result in taxonomies than ontologies. Instead, a consideration of further types of relationships such as *part-of* relations seems to be largely disregarded. In contrast to that, the ontology learning approach proposed by [32] deal with enriching the retrieved relationships with additional attributes, which originate from the knowledge sources. Thereby, [32] rely on annotations by means of the extrapolation of contextual relationships. A closer inspection of the different methodologies (statistics vs. NLP) with regard to a historical dimension shows that the different ontology learning approaches evolve over time starting with the identification of rather simple ontology components to the more advanced extraction of complex ontology components from the underlying knowledge sources.

The majority of the reviewed ontology learning approaches exhibit a lack in providing explicit information about the performance, i.e. concrete performance measures. Instead, the authors report on the target applications and the use of the proposed ontology learning approach rather in terms of a proof of concept. In this context, literature argues that an objective evaluation of the ontology learning approaches introduced by [21],[27], [30], and [32] is hardly feasible. In contrast to that, the remaining ontology learning approaches provide more detailed information about the performance results when evaluating their approaches. For instance, [25] evaluate their algorithm with regard to different levels of granularity. That is, [25] provide explicit information about how the algorithm performs on word disambiguation with the generated topic signatures. The results show that the algorithm performs best with respect to a coarse level of granularity. In contrast to that, a more fine-grained level of granularity leads to a drastic decline in performance. Further, [26] report on similar results. Whereas the enrichment of some classes performs well, i.e. propositions are added to classes; other classes are not subject of enrichment at all. This difference might be due to the greater potential of the distributional meaning of some classes, e.g., medical doctor is too generic to provide a basis for achieving an adequate degree of quality. Similarly, [28] provide statistical performance measurements with regard to the application of their clustering technique. Thereby, the evaluation results show that the proposed ontology learning approach achieves a slightly higher (approx. 1%) degree of performance with respect to the retrieved classes and the precision than comparable approaches operating on two different domains. A further advantage of this approach is that the classes provide some additional description, which supports the users in better understanding the retrieved ontology. Moreover, [29] draw upon a pattern-based approach to obtain a basic ontological structure and, in this context, reached 76.29 % of correctly discovered relations. In addition to that, almost 50% of all dictionary entries are correctly imported into the envisioned ontology. At last, [31] compare different techniques for signature creation. The results show that only a combination of different signature methodologies for signature creation generates adequate results in terms of accuracy.

These observations allow for drawing the following conclusions. Some of the reviewed ontology learning approaches operate (fully) automated and unsupervised. The remaining ontology learning approaches generate basic classes and provide relations between these classes. However, there is still a demand for manual intervention to complete the ontologies with regard to the specific requirements of the target applications. Manual interventions typically concerns the (pre-)selection of the knowledge sources or the manual evaluation of the retrieved ontology components, e.g., by a domain expert. This need for manual intervention aggravates a usage of the ontology learning approaches on a larger scale (e.g., inter-organizational Knowledge Management) because there is still considerable time, costs, and efforts required. This implies that automated and integrated activities that perform quality controls allow the application of ontology learning approaches in an up-scaling context such as Knowledge Management.

Based on that, this paper synthesizes the review results and their discussion by means of formulating eight hypotheses in the form of research questions. These eight research questions might provide a foundation to further develop ontology learning approaches from unstructured knowledge sources for advancing Knowledge Management.

Hypothesis 1: How ontology learning approaches can support the knowledge management core processes most effectively?

Hypothesis 2: What ontology learning methodologies (e.g., statistics, natural language processing) and which combinations of them are most effective for Knowledge Management?

Hypothesis 3: What ontology learning techniques and which combinations of them are most effective for Knowledge Management?

Hypothesis 4: What degree of automation of ontology learning is most effective for Knowledge Management?

Hypothesis 5: What types of knowledge sources are most effective for reuse in Knowledge Management?

Hypothesis 6: What ontology components are most effective for Knowledge Management?

Hypothesis 7: What evaluation techniques and metrics are most effective for assessing the adoption of ontology learning for Knowledge Management?

Hypothesis 8: How to empirically study the actual use and the realized benefits of ontology learning in Knowledge Management?

This set of hypotheses suggest research to not only focus on technical advances in ontology learning from unstructured knowledge sources for Knowledge Management but also explicitly include empirical research to study behavioral issues in terms of the impacts of current ontology learning approaches to further enhance the understanding, and, thus, fertilize future research in this area.

5 Conclusion

The objective of this paper was to give an account of the current state-of-the-art in order to contribute to an enhanced understanding of ontology learning from unstructured knowledge sources for Knowledge Management. On the basis of a review strategy, this paper identified nine ontology learning approaches. Four of these approaches primarily draw upon statistics whereas the remaining five approaches rely on natural language processing. To analyze the nine ontology learning approaches, this paper applies a classification framework, which consists of six criteria. These six criteria stem from literature and, thus, reflect descriptive constructs of the domain of ontology learning from a Knowledge Management point of view.

A literature review for the broad and fine-grained category of ontology learning from unstructured knowledge sources for Knowledge Management is a difficult task because of the large amount and diversity of background knowledge needed for studying, classifying, and comparing these ontology learning approaches. Therefore, the first shortcoming of this research is the authors' limited knowledge in presenting an overall picture of this subject. Secondly, some ontology learning approaches were not included in this literature survey (due the survey's purpose and scope). Third, the classification framework provides a set of six constructs that represent key descriptive constructs of ontology learning from a Knowledge Management point of view. The classification framework could be further extended and detailed with respect to ontology learning as well as Knowledge Management.

The results of the review provides evidence that research on ontology learning from unstructured knowledge sources has the potential to significantly enhance the current state-of-the-art in Knowledge Management. There is still a large gap between the multifaceted nature and the advancements in the core discipline of ontology learning and the actual use in Knowledge Management. Therefore, this paper proposes to extend the classification framework with additional criteria dealing with both issues of ontology learning and Knowledge Management to be able to elaborate in a more precise and fined-grained way on the potentials of ontology learning from unstructured knowledge sources for Knowledge Management. Moreover, this paper suggests that future research should be focused on the usefulness and efficacy of ontology learning approaches from unstructured knowledge sources for which Knowledge Management is a well-suited example because of rather weak dependency of specific domains or industries and its high relevance to practice.

Acknowledgement. The work presented in this paper was partly funded by the German Federal Ministry of Education and Research under the project InterLogGrid (BMBF 01IG09010E).

References

1. Staab, S.: Wissensmanagement mit Ontologien und Metadaten. Informatik-Spektrum 25, 194–202 (2002)
2. Staab, S., Schnurr, H., Studer, R., Sure, Y.: Knowledge Processes and Ontologies. IEEE Intelligent Systems 16, 26–34 (2001)
3. McComb, D.: Semantics in business systems. Morgan Kaufman, Massachusetts (2004)
4. Hazman, M., El-Beltagy, S.R., Rafea, S.: A Survey of Ontology Learning Approaches. International Journal of Computer Applications 20, 36–43 (2011)
5. Shamsfard, M., Barforoush, A.: The state of the art in ontology learning: a framework for comparison. The Knowledge Engineering Review 18, 293–316 (2003)
6. Biemann, C.: Ontology Learning from Text: A Survey of Methods. LDV Forum 20, 75–93 (2005)
7. Alavi, M., Leidner, D.E.: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. MIS Quarterly 25, 107–136 (2001)
8. Probst, G., Raub, S., Romhardt, K.: Wissen managen. Gabler Verlag, Wiesbaden (1998)
9. Smith, B.: Ontology. In: Floridi, L. (ed.) Blackwell Guide to the Philosophy of Computing and Information Blackwell 2003, pp. 155–166. Blackwell, Malden (2003)

10. Gruber, T.R.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human-Computer Studies 43, 907–928 (1995)
11. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5, 199–220 (1993)
12. Borst, W.: Construction of Engineering Ontologies for Knowledge Sharing and Reuse, PhD Thesis, University of Enschede (1997)
13. Studer, R., Benjamins, R., Fensel, D.: Knowledge Engineering: Principles and Methods. Data and Knowledge Engineering 25, 161–197 (1998)
14. Motta, E.: Reusable Components for Knowledge Modeling. In: Case Studies in Parametric Design. IOS Press, Amsterdam (1999)
15. Guarino, N.: Formal Ontology and Information Systems. In: Proceedings of the First International Conference on Formal Ontology in Information Systems, Trento, Italy (1998)
16. Uschold, M.: Building ontologies: Towards a unified methodology. In: Proceedings of the 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems, Expert Systems 1996, Cambridge, UK (1996)
17. Gómez-Pérez, A., Fernández-López, M., Corcho, O.: Ontological Engineering with examples from the area of Knowledge Management, e-Commerce, and Semantic Web. Springer, London (2004)
18. Corcho, O., Fernández-López, M., Gómez-Pérez, A.: Methodologies, tools and languages for building ontologies. Where is their meeting point? Data and Knowledge Engineering 46, 41–64 (2003)
19. Navigli, R., Velardi, P., Faralli, S.: A Graph-Based Algorithm for Inducing Lexical Taxonomies from Scratch. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Spain, pp. 1872–1878 (2011)
20. Sabou, M., Wroe, C., Goble, C., Mishne, G.: Learning domain ontologies for Web service descriptions: an experiment in bioinformatics. In: Proceedings of the 14th International Conference on World Wide Web, Chiba, Japan, pp. 190–198 (2005)
21. Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: Proceedings of the 14th Conference on Computational Linguistics, Stroudsburg, PA, USA, pp. 539–545 (1992)
22. Cimiano, P., Mädche, A., Staab, S., Völker, J.: Ontology Learning. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies 2009, pp. 245–267. Springer, Heidelberg (2009)
23. Buitelaar, P., Cimiano, P., Magnini, B.: Ontology Learning from Text: An Overview. In: Magnini, B., Buitelaar, P., Cimiano, P. (eds.) Ontology Learning from Text: Methods, Evaluation, and Applications, pp. 1–13. IOS Press, Amsterdam (2005)
24. Mädche, A., Staab, S.: Learning Ontologies for the Semantic Web. IEEE Intelligent Systems 16, 72–79 (2001)
25. Agirre, E., Ansa, O., Hovy, E., Martinez, D.: Enriching very large ontologies using the WWW. In: Proceedings of the ECAI Workshop on Ontology Learning, Berlin, Germany (2000)
26. Faatz, A., Steinmetz, R.: Ontology Enrichment with Texts from the WWW. In: Proceedings of the 13th European Conference on Machine Learning, Helsinki, Finland (2002)
27. Sanchez, D., Moreno, A.: Creating Ontologies form Web Documents. Recent Advances in Artificial Intelligence Research and Development 113, 11–18 (2004)
28. Cimiano, P., Hotho, A., Staab, S.: Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. Journal of Artificial Intelligence Research 24, 305–339 (2005)
29. Kietz, M., Volz, R., Mädche, A.: A method for semi-automatic ontology acquisition from a corporate intranet. In: Proceedings of EKAW 2000 Workshop "Ontologies and Text", Juan-Les-Pins, France (2000)

30. Gupta, K.M., Aha, D.W., Marsh, E., Maney, T.: An Architecture for Engineering Sublanguages WordNets. In: Proceedings of the First International Conference on Global WordNet, pp. 21–25 (2002)
31. Alfonseca, E., Manandhar, S.: Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 1–7. Springer, Heidelberg (2002)
32. Narr, S., DeLuca, E.W., Albayrak, S.: Extracting semantic annotations from Twitter. In: Proceedings of the Fourth Workshop on Exploiting Semantic Annotations in Information Retrieval, ESAIR 2011, pp. 15–16 (2011)
33. Harris, Z.S.: Mathematical Structures of Language. Wiley, New York (1968)
34. Kullback, S., Leibler, R.A.: On Information and Sufficiency. The Annals of Mathematical Statistics 22, 79–86 (1951)
35. Schütze, H.: Foundations of statistical natural language processing. The MIT Press, Massachusetts (1999)
36. Hahn, U., Marko, K.G.: Joint Knowledge Capture for Grammars and ontologies. In: Proceedings of the 1st International Conference on Knowledge Capture, pp. 68–76 (2001)