14 The life cycle of knowledge

Alessandro Lenci

14.1 Introduction

Due to its *prima facie* abstract character, we are often lead to conceive *knowledge* as it were a changeless entity inhabiting an extra-time realm, much alike Plato's World of Ideas. As a consequence, we tend to forget or to underestimate the crucial fact that knowledge has its own life cycle too, not very differently from anything else in the natural Universe. Knowledge is created. It changes through time. It reproduces itself, generating new knowledge. Knowledge dies too, as proved by the fact that our knowledge about artifacts, practices, people or places can get lost or disappear forever. In technical domains (such as for instance biology, medicine, computer science, but also agriculture, as claimed by Kawtrakul and Imsombut (Chapter 17 this volume)), the rate of knowledge change can be very high, with concepts becoming 'obsolete' because of new technological advances, which in turn may induce complex reorganizations or expansions of the knowledge system. It goes without saying that similar processes apply to ontologies as well, insofar as they are formal systems aiming at representing a certain body of knowledge, thereby being closely tied to its destiny.

The dynamics of knowledge and of the ontologies that represent it depend on its *contexts of use*. Knowledge is created or acquired for some purpose, i.e. to be used as a tool to achieve a certain goal or to perform a particular task. Use also changes our knowledge about entities and processes and consequently leads us to revise our ontological systems. Moreover, the employ of some body of knowledge to perform a task may produce new knowledge that has to be added to our ontologies, possibly resulting in a major revision of their structure, if some

breakthrough in the knowledge system has occurred. Actually, this is usually the main or ultimate purpose for us to carry out the task itself.

This book deals with ontological and lexical knowledge resources. The *ontolex* interface (See chapter 10 this volume) is an attempt to answer to the increasing need of modeling the complex interrelationship between lexicons and ontologies, which are more and more assuming the form of rich *ontolexical* resources. I will use this term to stress the importance of investigating the substantial areas of overlapping between ontologies and lexicons, as a step toward a better understanding of their individual characters. In this chapter, I will focus on the interaction between *ontolexical* knowledge systems and a specific context of use that shapes their dynamics, i.e. *natural language processing* (NLP). Therefore, methods, tools and applications for NLP are here analyzed under the perspective of how they affect the life cycle of knowledge.

There is a bidirectional link between NLP and *ontolexical* resources. First of all, NLP tools and applications are intensive *knowledge users*, i.e. they have to access large amounts of different types of knowledge to carry out the tasks they are designed for. This raises the issue of how the particular goals of NLP systems act as constraints on ontological systems, so that the type and the organization of knowledge can optimally comply with the needs of NLP tasks. Secondly, NLP tools and applications are also powerful *knowledge creators*, i.e. they can be used to create and modify *ontolexical* resources, for instance by allowing the bits of knowledge to be represented in such resources to be carved out of the linguistic structures that encode them in natural language texts. Despite their *prima facie* differences, the two roles of NLP systems as knowledge users and knowledge creators must be conceived as deeply interrelated one to the other, actually being two sides of the same coin. This integrated view is not an optional one, but rather it represents the essential condition for NLP systems on the one hand to exploit at their best the incredible potential offered by large-scale *ontolexical* resources, and on the other hand to boost and enhance the process itself of knowledge creation.

14.2 Using ontolexical knowledge in NLP

From Information Extraction to Question Answering, the final goal of most NLP systems and applications is to access the information content of texts through the interpretation of their linguistic structures. To carry out their tasks, NLP systems need to know the relevant pieces of knowledge to be identified in texts, as well as how this knowledge is encoded in linguistic expressions. The role of *ontolexical* resources is to provide NLP systems with these two crucial types of information. Besides, there is a further important factor that assigns a key role to *ontolexical* knowledge in NLP, a factor that is not directly dependent on the

specific system goals, but rather derives from the core principles of language. In fact, much of the constraints natural language grammars obey to are *lexicalized*. That is to say, many aspects of complex linguistic structures depend on the properties of the lexical items that compose them. The *paradigmatic* properties of lexical items – i.e. the classes to which they belong in virtue of their morphosyntactic and semantic properties – constrain their possible *syntagmatic* distributional properties – i.e. the range of linguistic contexts in which they can appear. Thus, *ontolexical* knowledge can not but playing a prominent role in most NLP architectures. Although there are approaches that dispense with this type of information altogether, the addition of *ontolexical* information usually proves to be a necessary step to enhance the system performances in tackling the challenges set by natural language.

In the sections below, I will focus on four major use contexts for *ontolexical* knowledge in NLP: *semantic typing*, *semantic similarity and relatedness*, *inference*, and *argument structure*. They represent some of the most important ways in which systems typically employ the information in *ontolexical* resources to achieve their specific applicative goals. While applications often change quite rapidly – following the waves of market needs or of technological developments – these use contexts provide us with general vantage points from which to observe the role of different aspects of semantic information in NLP, as well as the main problems and challenges that knowledge intensive processing of natural language must face.

14.2.1 Semantic typing

The primary aim of *ontolexical* resources is to characterize the *semantic types* of linguistic expressions, i.e. the classes to which linguistic expressions belong in virtue of their meanings. Types can be regarded as formal, symbolic ways of identifying the concepts expressed by linguistic expressions. To the extent that meanings are related to entities in the world, semantic types also correspond to the categories of entities referred to by linguistic items. By assigning a lexical item to a semantic type we thus characterize and focus on specific aspects of its semantic space. For instance, the type AIRPLANE can be used to explicitly represent one of the senses expressed by the word *'plane'*.

Ontologies and lexicons represent key resources for *semantic typing*, i.e. the process of automatically identifying the semantic types of linguistic expressions in texts. As Pustejovsky et al. (2002) rightly claim, semantic typing is the backbone and prerequisite of most NLP applications to achieve content-based access to texts. For example, the goal of Information Extraction is to identify relevant facts from texts. Facts are typically defined by events involving a certain number of entities belonging to different categories (e.g. humans, locations,

proteins, vehicles, etc.). The instances of the semantic categories relevant for the application target domain must be identified in texts, and this amounts to assigning the proper semantic type to their linguistic descriptors. Take the following pair of sentences: *The new medicine is highly effective against pneumonia. This illness can still be quite dangerous*. An Information Extraction system operating in the bio-medical domain might need to understand that the word *'pneumonia'* belongs to the semantic type DISEASE and that therefore the first sentence actually mentions a fact or event concerning this category of entities. Similarly, type identification is also critical to resolve the anaphorical link between the noun phrase *this illness* in the latter sentence and *'pneumonia'* in the former one.

Besides the semantic typing of entities, the problem of identifying the relations linking these entities is gaining increasing attention (Nastase and Szpakowicz, 2003; Girju et al., 2005; Turney, 2006). Semantic relations can be explicitly encoded by lexical items, such as verbs or relational nouns, but they can also be implicitly expressed by linguistic constructions. For instance, the proper interpretation of noun compounds require the understanding of the specific relation holding between their constituent words, such as **material** (*'apple pie'*), **location** (*'lung cancer'*) or **meronymy** (*'door handle'*). Conversely, the semantic relation of **causation** , besides being explicitly expressed by lexical items such as *'cause'* or *'provoke'*, is also implicitly encoded by compound nouns such as *'food infection'* or *'flu virus'*. Most of these relations are represented in *ontolexicons*, which are therefore often used as key knowledge resources for automatic semantic relation identification.

As it clearly emerges from these few examples, semantic typing is actually a very complex task, which in turn presents a high spectrum of possible variations, depending on the applicative and the domain specific needs. In any case, it crucially relies on the identification of the semantic potential of lexical items, and more precisely on the possibility to characterize the way linguistic constructions express a certain system of semantic categories and the relations among these categories. This is the reason why semantic typing represents a crucial use context for *ontolexical resources*.

14.2.2 Semantic similarity and relatedness

Natural language expressions can share different aspects of their meanings, that is to say they can be *semantically similar* at various degrees. Semantic similarity is a very loosely defined notion, actually forming a wide spectrum of variation spanning from the full– or near–synonymy of pairs such as *'king'–'monarch'*, to much wider associative links. Budanitsky and Hirst (2006) (following Resnik (1995)) distinguish semantic similar pairs such as *'horse'–'pony'* from semantic related ones, such as *'hot'–'cold'* or *'handle'–'door'*. While similar pairs contain

words referring to entities that share a certain number of salient 'features' (e.g. shape, position in a taxonomy, functionality, etc.), related pairs are formed by words that are connected by some type of semantic or associative relation – such as antonymy, meronymy, frequent co-occurrence – without being necessarily similar themselves (e.g. *'handle'–'door'*). Parallel distinctions are typically assumed in the psycholinguistic literature, where the notion of semantic similarity plays a crucial role in the explanation of phenomena such as priming effects (Moss et al., 1995).

A huge literature exists in linguistics, philosophy and cognitive science devoted to the many conundrums hidden behind such a *prima facie* natural aspect of semantic organization. The problem is that, while the fact that *'dog'* is more semantically similar to *'cat'* than to *'car'* appears to us as more or less incontrovertible, turning this intuition into effective and formal criteria to determine the degree of semantic similarity between two words is extremely hard. Yet, measuring the semantic similarity (or relatedness) between words has a key importance in any applicative context, such as Word Sense Disambiguation, Information Extraction and Retrieval, Machine Translation, etc.

Ontolexical systems play an important role in computing semantic similarity (relatedness). Besides the fact that most of these resources explicitly contain lists of synonymous terms (like the synsets in WordNet), groups of similar words can be explicitly represented by assigning them to the same semantic type. For instance, *'airplane'*, *'boat'*, *'car'* and *'bus'* can all be assigned to the type VEHICLE, thereby making explicit the fact that they belong to the same paradigmatic class as determined by key features of the meaning they share (e.g. they can move, are designed for transportation, can be driven, etc.). If conversely *'cat'* and *'dog'* are assigned to the type ANIMAL, the 'dissimilarity' between *'dog'* and *'car'* immediately descends from their belonging to different semantic types. Even more crucially, it is the structure itself of the ontology that can be used to compute the degree of semantic similarity (relatedness) between words. In fact, semantic similarity (dissimilarity) between two words can be formally defined in terms of the closeness (distance) of their types in the semantic space defined by the ontology. Various types of metrics have been proposed in the literature that exploit the path (and types) of relations connecting two concepts to measure their semantic proximity. Budanitsky and Hirst (2006) provides an interesting survey and evaluation of different measures of semantic relatedness based on the topology of relations in WordNet.

14.2.3 Inference

The main function of the human conceptual system is to provide the basis for drawing inferences about the entities that belong to a certain category (Murphy,

2002). In a parallel fashion, this inferential ability is a characteristic property of our lexical competence: knowing the meaning of a word is also being able to draw some inferences from it (Marconi, 1997). For instance, if we understand the meaning of the sentence *Tweety is a bird*, we also infer that Tweety is an animal and that it is very likely to fly. Similarly, the meaning of the word *'kill'* in the sentence *The man killed the gorp* allows us to infer that the *'gorp'* – whatever this entity might be – was very likely to be a living being before the occurrence of this event, and became dead afterward. Inferences differ with respect to their type and strength, some of them being logical entailments (as in the case of a cat being an animal), others having instead just a probabilistic value. In fact, not all birds fly, but only the most prototypical ones. In either case, inferences depend on the properties and on the organization of our system of concepts and meanings.

Ontolexical resources represent the most direct way to explicitly capture the inferential relations between concepts and semantic types. Usually, they are not inferential system *per se*, but rather they are representational resources on which such systems can be defined. The definition of the classes of an ontology and the network of relations connecting them are the basis to define their inferential properties. Similarly to the cognitive processes of categorization and concept formation, semantic types are designed at various levels of generality, abstracting away from specific features of meanings. Actually, semantic types usually form chains of conceptual classes ordered by subsumption relations: AIRPLANE, FLYING_VEHICLE, VEHICLE INANIMATE_OBJECT, CONCRETE_OBJECT. Such chains allow systems to draw inferences that are crucial in many applicative contexts in NLP, such as Information Extraction, anaphora resolution, textual entailment recognition (Dagan et al., 2006), etc. Consider for instance the following sequence of sentences: *The bus suddenly stopped along the road and the passengers went out of the vehicle. The engine was broken.* Capturing its information content requires NLP systems to resolve some cases of "bridging" anaphora, i.e. the referential phenomenon occurring when the referent of a linguistic expression can be determined only by recovering a meaningful implicit relation with an already mentioned entity or event. The co-reference between *'vehicle'* and *'bus'* can be resolved if the hyperonymic relation holding between these nouns is known to the system. Similarly, the availability in a computational lexicon of the information that engines are parts of buses, can lead the system to infer the obvious fact that the broken engine belongs to the bus mentioned in the first sentence.

14.2.4 Argument structure

One of the most common use contexts for *ontolexical* resources in theoretical and computational linguistics is to specify the combinatorial constraints of lexical

items. Some *ontolexicons* provide explicit representational devices of argument structure properties, such as number and semantic types of arguments, semantic roles, argument alternations and realizations (Levin and Hovav, 2005), etc. This is for instance the case of FrameNet (Baker et al., 2003), Omega (cf. chapter 15 in this volume), VerbNet (Kipper-Schuler, 2005), SIMPLE (Lenci et al., 2000), among the others. It is also a common practice to start from a given ontology of types and then to try to use its conceptual atoms to specify the *selectional preferences* of predicative expressions. For instance, although WordNet itself does not encode argument structure properties, it has been widely used as a source of semantic types for argument semantic specification (Resnik, 1996). Light and Greiff (2002) and Brockmann and Lapata (2003) provide interesting surveys and evaluations of various WordNet based approaches to selectional preferences. Generally, the structure of the ontology is exploited in order to identify the suitable level of semantic abstraction of the arguments that can typically occur in a certain predicate role (e.g. as direct objects of *'eat'*, or as subjects of *'drive'*). Given the notorious difficulties of defining the predicate selectional preferences as sets of necessary and sufficient conditions, probabilistic models are typically used to establish the proper mapping, i.e. to determine the type or types that best capture the combinatorial constraints of predicates.

It goes without saying that the role of ontolexical resources in providing suitable representations for predicative structures is of paramount importance for a large number of tasks in NLP. Selectional preferences can act as key constraints for parsing, Question Answering, and relation extraction. Moreover, event identification in texts also requires access to information about the semantic roles expressed by predicates. Semantic role labeling (Gildea and Jurafsky, 2002; Erk and Padó, 2006b) exploits resources such as FrameNet that provide the information about predicate argument structures, argument semantic roles, and the inferential relations between these roles (e.g. that a driver is a type of agent).

14.2.5 The challenges of the ontolex interface

So far so good. *Ontolexical* resources appear to be undoubtedly important components in knowledge-intensive NLP applications involving tasks that crucially depend on knowledge about the structure and organization of a conceptual domain, and on the semantic content of the lexical and grammatical constructions describing this domain. *Ontolexical* systems can fulfill this role to the extent that they are able to provide suitable formal characterizations of the repertoire of semantic types and of the mapping between the language system and the conceptual system. But, what are the challenges to achieve these goals?

The main problem is that the interface between language and concepts that *ontolexicons* purport to represent is notoriously highly complex. The principles

governing this interface are still under many respects obscure and to a large extent defy precise formalizations. In general, no naive mapping between lexical and conceptual systems can pretend to capture the order of complexity shown by the semantic behavior of natural language expressions. Polysemy, metaphor, metonymy and vagueness are only some examples of the rich semantic phenomenology through which this complexity manifests ubiquously in language. These phenomena are *systematic*, in the sense that they present specific regularities within a language and across languages (Pustejovsky, 1995). At the same time, they represent also *systemic* features of natural language, since they are inherent properties of its semantic organization and of the way the mapping between the conceptual and language systems have become established.

These pervasive semantic phenomena point toward a non-naive relationship between *concepts* - as representations of categories of entities - and *meanings* - as the semantic content of linguistic expressions. Most of the theoretical and computational literature in semantics (this chapter included) usually tends to treat these two terms as essentially interchangeable, with meanings being regarded as concepts mapped on or linked to linguistic symbols that are conventionally used to communicate them. Actually, this equation is more or less explicitly assumed in many computational lexicons (e.g. WordNet), whose word senses descriptions are often interpreted as concepts of an ontology.[1] Its widespread use notwithstanding, the meaning–concept equation is however not granted at all. Indeed, in recent psychological research on human semantic representation and categorization systems, there is rich evidence supporting the view that these two notions should be kept well distinguished (Murphy, 2002; Vigliocco and Vinson, 2007). This, obviously, does not mean to deny that concepts and meanings are related, but rather that this relation can not be assumed to be one of straightforward "ontological"F identity. In fact, if this assumption is dropped, the notion itself of *ontolexical* interface gains much more relevance as the place at which the complex interplay between conceptual systems and the semantics of natural language can be represented and investigated.

A second issue raised by semantic phenomenology is the relationship between the meanings of lexical expressions as captured by *ontolexical* resources and their interpretation in context. *Ontolexical* resources are generally systems of semantics types which are defined and characterized more or less independently of the typical linguistic contexts in which these types are used. The crucial issue is to understand the extent to which context enters into the semantic constitution of linguistic expressions. In fact, lexical meaning is to a large extent a context-sensitive reality, and phenomena like polysemy or metonymy should be more

[1] There are also exceptions. An example is provided by the *Omega Ontology* (chapter 15 this volume), which includes an explicit distinction between the level of word sense description and the level of conceptual representation.

properly modeled as the results of sense generation processes in context (Pustejovsky, 1995). Lexical expressions have a semantic potential that gets realized as specific senses or interpretations when they combine with other linguistic expressions in syntagmatic contexts. It is worth emphasizing that the importance of these issues greatly exceeds their centrality for theoretical semantic research. Given the aim of ontolexical systems at being effective knowledge resources for NLP systems, they can not but face the challenges set by the different creative ways in which lexical items are used in texts. Tackling these phenomena raises the key question of whether they should be accounted at the representational level, i.e. within the lexicon or the ontology, or rather at the processing level, i.e. by those systems that will use *ontolexical* knowledge within the larger perspective of text understanding. The answer to this notorious dilemma deeply affects the structure of *ontolexical* resources themselves, as well as the way they can be used in NLP tasks. For instance, wiring too many polysemous or metaphorical uses in the lexicon typically results into very granular sense distinctions. These will in turn negatively impact on systems that need to map such fine-grained sense distinctions on texts, e.g. for Word Sense Disambiguation or Machine Translation. Therefore, enhancing the usability of lexical semantic resources for NLP tasks necessary requires a better understanding of the theoretical principles governing the *ontolex* interface. Actually, its domain should not be limited to the characterization of binary relations between concepts and lexical items, but should instead cover the threefold interaction between the conceptual system, lexical expressions and the linguistic contexts that shape and modulate their senses.

14.3 Creating ontolexical knowledge with NLP

Every time we introduce a new item in a ontology we perform "an act of creation" (Hovy, 2005). Typically, this demiurgic experience is carried out by a domain expert, who builds the ontology either directly or indirectly, in the latter case by providing another 'ontologizer' with the necessary information about the domain structure. In either case, the human expert is supposed to be the most reliable knowledge source, from which the various components of an ontology can be made explicit and formalized. The key role of human expertise notwithstanding, another not less crucial knowledge source for ontology building is represented by *natural language texts*. Indeed, documents – from Wikipedia to scientific papers and technical reports – are the primary repository of the knowledge of a certain community. Therefore, they can be mined to identify the knowledge items most relevant to characterize a particular domain, and use them to feed the ontology creation process.

The challenge in using document sources for ontology development is obviously how to carve the formal structure of the conceptual system out of the implicit and informal ways in which knowledge is expressed and encoded in texts. The role of NLP methods and tools is exactly to help to bridge this gap, by extracting the relevant pieces of knowledge from texts through various levels of processing and analysis of their linguistic structure. The use of NLP – in combination with machine learning, and AI-derived methods – to acquire knowledge from texts in support to ontology development is now commonly referred to (especially in the Semantic Web community) as *ontology learning*.[2] Indeed, ontology learning holds a high 'family resemblance' with the long-standing line of research in computational linguistics concerning (semi-)automatic acquisition of lexical information from texts (Manning and Schütze, 1999) in support to the development of computational lexical resources. Although there are some reasons to keep these two fields apart (Buitelaar et al., 2005), the existence of a strong commonality of methods and intents is undeniable. Indeed, many popular techniques for ontology learning were originally born and developed in the context of lexical acquisition. The possible complementarity between ontology learning and lexical acquisition rather lies in a difference of emphasis with respect to their goals. While the purpose of ontology learning is to support the development of conceptual resources, lexical acquisition is generally more oriented toward the text-driven acquisition of specific linguistic properties of lexical items (e.g. subcategorization patterns, selectional preferences, synonymy relations, etc.).

In the context of *ontolexical* resources – with their strong interaction and interplay between ontological and lexical knowledge – clearly the overlapping between ontology learning and lexical acquisition also increases. The boundaries between these two enterprises become so tenuous that the term of *ontolexical learning* seems to be perfectly justified. What is worth emphasizing is that in both cases NLP directly enters into the life cycle of knowledge, by supporting the process of creation and growth of *ontolexical* resources. The latter is surely the phase in which the role of text-driven learning is most effective. In fact, although there are cases in which a whole ontology is bootstrapped from natural language sources, a much more common scenario is the one in which NLP techniques are used to extend and enrich an existing, human-made *ontolexical* resource. For instance, there are countless works focusing on the (semi)-automatic extension of WordNet through lexical information automatically harvested from corpora. Pustejovsky et al. (2002) apply NLP and statistical techniques to extend and adapt UMLS, the most important knowledge organization system in the medical domain. Similar techniques are also used by Kawtrakul and Imsombut (chapter 17, this volume) for the maintenance of an ontology in the agricultural domain.

[2] Actually, ontology learning is a much broader field encompassing also knowledge acquisition from non-textual sources. Nevertheless, I will use this term only to refer to knowledge acquisition from texts.

The spectrum of solutions offered to the problem of ontolexical learning is incredibly wide, and the number of publications devoted to it huge. Rather than attempting an impossible as much as useless survey of existing approaches to create knowledge with NLP methods, I will here focus on three general questions: *what pieces of knowledge can we extract with NLP? from which sources? how is NLP used to extract ontolexical knowledge?* There is a fourth issue that is also worth touching: *What for?* That is to say, *what are the advantages of using NLP and learning methods for ontolexical building?* Apparently, there is a very simple and direct answer, which usually appears at the beginning of every paper on this topic, i.e. because it is convenient, since it makes the process of *ontolexical* development easier and faster. However, we will see that this is not the only rationale, and there are actually more theoretical reasons that suggest that extracting *ontolexical information* from text data may actually enhance its quality and usability.

14.3.1 Which ontolexical information can be extracted with NLP?

Ontologies are complex entities that contain different types of components (cf. chapter 10 this volume): classes representing categories of objects, properties and relations, links to the linguistic constructions that express these conceptual entities in a given language, cross-lingual links, etc. Actually, this whole spectrum of entities can be the target of NLP-based acquisition processes. Buitelaar et al. (2005) propose to arrange the possible targets of knowledge acquisition in what they refer to as the *ontology learning layer cake*: from the bottom to the top, this includes *terms*, *synonyms*, *concepts*, *concepts hierarchies*, *relations*, and *rules*. As it is clear from this list, the layers differ with respect to the increasing degree of abstraction from the linguistic surface, and consequently with respect to the complexity of the learning task itself. Extracting the relevant domain terminology from a text collection is a crucial step within ontology learning and is now mature for real-scale applications (Frantzi and Ananiadou, 1999; Jacquemin, 2001). Synonymy detection methods have achieved impressive results, and their performance is now approaching human-like performance (Lin, 1998b; Rapp, 2003). Much effort is also devoted to the text-driven identification of taxonomical and other non-hierarchical relations (Hearst, 1992; Cimiano et al., 2005; Pantel and Pennacchiotti, 2006), although further research is needed to improve the accuracy of relation learning.[3]

Besides the entities mentioned in the "layer cake", other important pieces of knowledge that can be acquired automatically from texts are the instances of the ontology classes. This issue is usually considered to lie outside the specific

[3]For further references on these issues cf. Buitelaar et al. (2005).

field of ontology learning, and receives the name of *ontology population*. Yet, it is surely an important part of the general ontology development cycle: for instance, acquiring information about which entities in a domain are instances of a particular ontology class, is an important condition to enhance the usability of the ontology for various tasks or applications (cf. chapter 16 in this volume, for the case of Question Answering). Interestingly, Nédellec and . (2005) point out the connections between Information Extraction - as the task of extracting factual information about events and entities - and ontology population. This an important case of a virtuous circle in which an NLP core task such as Information Extraction at the same type can play the role of *ontolexical* knowledge user and developer.

Moving toward the linguistic side of *ontolexical* resources, the range of information types that are targeted by text-driven acquisition processes is equally extremely wide. Besides term extraction and synonymy detection that are shared with the ontology learning enterprise, we can mention the acquisition of subcategorization frames (Korhonen, 2002), predicate selectional preferences (McCarthy, 2001), lexicalized concept properties (Almuhareb and Poesio, 2004; Cimiano and Wenderoth, 2007), etc. Again the state-of-the-art performances are strictly correlated with the type of targeted lexical information. In general, however, it is safe to assume that automatically extracted information is constantly gaining centrality and importance for the development and extension of *ontolexical* resources.

14.3.2 Which text sources can be used?

Until now we have generally talked of text-driven knowledge extraction, but actually an important parameter in *ontolexical* learning methods concerns the type of natural language source. A major divide exists between knowledge extraction from *semi-structured texts* such as thesauri, glossaries and machine readable dictionaries, and knowledge extraction from *text corpora*. In the former case, the input is represented by texts that are already designed to act as knowledge and lexical resources, although their structure is typically not a formal one and usually addressed to a human user. For instance, a long-standing line of research in computational linguistics has applied NLP techniques to convert dictionary definitions into structured semantic entities to populate computational lexicons (cf. for instance the ACQUILEX projects). The advantage of using existing human-oriented knowledge resources is the possibility to exploit their partial structure (e.g. the conceptual categories of a thesaurus or the sense distinctions in a dictionary) to spell-out the domain conceptual organization. In fact, the thesauri or glossaries already available in many technical domains represent major repositories of the knowledge shared by a community, and therefore provide key input to determine the relevant domain concepts and structure. On the other hand,

the use of these semi-structured text sources have shortcomings as well, the most important one being the fact that they are themselves limited and biased by the fact of being originally designed for human users. For instance, a dictionary entry for a word may give useful information for a human reader to understand bits of its meaning, but at the same time fail to provide key semantic properties necessary for an application to process that same word in a certain NLP task.

To overcome the limitations of semi-structured lexical resources, most approaches to *ontolexical* learning use *text corpora*.[4] The basic assumption is that the concepts relevant to organize a particular domain of knowledge can be extracted from texts representative of that domain. Although the fact that the relevant pieces of knowledge are only implicitly encoded in texts provides a high challenge for NLP methods, corpus processing is surely the most promising source for ontology population, extension and maintenance. Since ontology learning is mostly directed toward the vertical enrichment of domain ontologies, specialized corpora represent the preferred data source. For instance, the Medline collection of medical abstract can be an endless knowledge mine to refine and extend medical ontologies (Pustejovsky et al., 2002). On the other hand, large scale, open domain corpora are used as well. Large corpora are in fact useful to address or limit the negative effect of data sparseness, and many approaches now regard the Web itself as an important resource to extract semantic information. The applications of NLP methods to on-line encyclopedia such as Wikipedia could represent an interesting compromise between the use of semi-structured knowledge source and corpus processing for ontology learning.

14.3.3 How to use NLP to extract ontolexical knowledge?

Ontolexical learning is typically carried out through some mixture of text analysis with NLP tools – from lemmatization and PoS tagging to different forms of shallow and deep parsing – together with machine learning or statistical methods to identify and weigh the extracted pieces of knowledge. The particular type of linguistic processing and statistical method provide the main parameters of variation among existing approaches to *ontolexical learning*. Within this large spectrum, one major family of algorithms is based on the *a priori* identification of *linguistic patterns* univocally associated with particular types of knowledge or semantic relations. This trend of research has been opened by the seminal work of (Hearst, 1992), who used automatically extracted patterns like *such*

[4]Hybrid approaches also exist, that use combinations of text corpora and structured knowledge resources for ontology learning. Cf. for instance Kawtrakul and Imsombut (chapter 17 in this volume).

NPh as NP1 or NP2 to identify pairs of concepts linked by hyperonymic or co-hyponymic relation. For instance, the identification in a corpus of the expression *such aircrafts like jets or helicopters* as an instance of the above pattern, would be taken as evidence that *'jet'* and *'helicopter'* are **hyponyms** of *'aircraft'*. This method includes a sort of weak supervision, since the linguist must decide *a priori* which patterns are associated with which semantic relations. The patterns can be more or less abstract depending on the type of text processing that is performed (e.g. tokenization, shallow parsing, etc.). They are then searched on a large corpus and then statistically filtered to reduce noise. The increasing popularity of pattern-based methods (Berland and Charniak (1999, Widdows and Dorow (2002, Almuhareb and Poesio (2004, Cimiano and Wenderoth (2007) among many others) is due to the fact that they are very promising in allowing the explicit 'typing' of the extracted knowledge, thereby facilitating its possible mapping onto existing ontologies. Yet, this strategy has also various shortcomings. First of all, it often runs into data-sparseness problem, since the relevant patterns are generally very rare. A common way to mitigate this problem is to use Web searches to have reliable statistics of linguistic patterns. Secondly, and most crucially, this approach works well only provided that we are able to identify easy-to-mine patterns, *univocally* associated with the target knowledge type. The problem is that this univocity is very rare in natural language, and most patterns are ambiguous or polysemous, since they encode very different types of semantic relations. For instance, the pattern *X has Y* can be used to expressed a meronymic relation (e.g. *a car has four wheels*), but also a possessive relation (e.g. *this man has a car*). This may result in high levels of noise, negatively impacting on the system precision.

Other methods for knowledge extraction instead adopt a fully unsupervised approach, and try to construct lexical semantic representations out of word statistical distributions in corpora. Rather than searching for the words that instantiate a number of pre-selected and (supposedly) semantically meaningful patterns, *semantic space models* represent a target word as a point in a n-dimensional vector space, constructed from the observed distributional patterns of co-occurrence of its neighboring words. Co-occurrence information is usually collected in a frequency matrix, where each row corresponds to a target word and each column represents its linguistic context. The assumption lying behind this type of semantic representation is the so–called "distributional hypothesis", i.e. that two words are semantically similar to the extent that they occur in similar contexts (Harris, 1968; Miller and Charles, 1991). Vectorial representations are used to estimate the semantic relatedness between two words on the grounds of their distance in the n-dimensional vector space. A huge spectrum of variation exists among these models, mostly due to the particular statistical and mathematical technique used to process the co-occurrence vectors, and to the definition of linguistic context. In *Latent Semantic Analysis* (Landauer and Dumais, 1997) the context is represented by a whole document in a collection and the word semantic similarity is computed in a reduced dimensionality space, obtained through the Singular Value

Decomposition of the original frequency matrix. Alternatively, the context can be provided by a window of n words surrounding the target word (cf. *Hyperspace Analogue to Language*, (Burgess and Lund, 1997).[5] Other approaches instead adopt a syntactically-enriched notion of context (Lin, 1998b; Padó and Lapata, 2007): i.e. two words are said to co-occur if they are linked by a certain syntactic relation (e.g. subject, modifier, etc.). This latter method is often claimed to be able to produce much more accurate semantic spaces, although a much heavier corpus pre-processing is required.

Semantic space models are very good in finding synonym pairs. For instance, Rapp (2003) reports 92.50% of accuracy in the synonym detection task carried out on the TOEFL dataset. Still, a major shortcoming of these methods is represented by the fact that their outcome is typically formed only by a quantitative assessment of the degree of semantic association between two words, with the type of relation remaining totally underspecified. Actually, the space of semantic neighbors of a target word can be highly heterogeneous, and besides synonyms it is typically populated by meronyms, co-hyponyms, or simply words belonging to the same semantic domain. Therefore, while the output of these methods can surely provide useful hints to evaluate the degree of semantic relatedness between two or more words, much work is still needed to carve actual semantic structure out of distributional spaces.

14.3.4 Why extracting ontolexical knowledge using NLP?

It is well-known that the process of developing ontologies and computational lexicons by hand is a very time– and money–consuming enterprise. Thus, the possibility offered by ontology learning methods to automatize parts of this process seems to be a promising way to overcome the so-called "knowledge acquisition bottleneck" (Maedche and Staab, 2004), i.e. the fact that *ontolexical* resources are terribly needed to perform new innovative steps in information technology, and yet they are also very slow and complex to develop. The possibility of speeding up this process thus surely counterbalances the fact that the extraction methods are far from being perfect and noise free. Moreover, knowledge acquisition is commonly regarded not as a stand-alone method for ontology expansion or population, but rather as a support to the unavoidable human intervention by domain experts. Within the context of the so-called *Balanced Cooperative Modeling* (Morik, 1993), text-driven knowledge extraction is just a phase in the ontology development cycle that must be complemented by human-made pruning

[5] But see also *Random Indexing* (Karlgren and Sahlgren, 2001), *Infomap* (Widdows, 2003), and *Incremental Semantic Analysis* (Baroni et al., 2007).

and refinements to integrate the acquired knowledge within existing knowledge resources.

Their importance notwithstanding, the practical needs of the *ontolexical* development process should not be the main rationale to pursue NLP–based approaches to knowledge acquisition. Actually, the use of text–driven learning methods appears as a necessary condition to grant the effective usability of *ontolexical* resources by NLP systems. Ontologies are nowadays used in many information processing contexts, most of them not directly concerned with natural language understanding. In these cases, documents can be important sources of knowledge for ontology development, but nevertheless still just complementary ones. Conversely, when we talk about using ontologies by NLP systems for the purposes of understanding and extracting information content encoded in natural language documents, then the situation is totally different. In fact, NLP needs ontolexical resources that are well "adapted" to the texts that they are going to process. (Pustejovsky et al., 2002) present various cases in which even a very rich and fine–grained domain ontology such as UMLS can not be effectively used for a NLP task such as anaphora resolution because of the very frequent type–mismatching. In fact, words that in texts are used co-referentially and as belonging to the same type may happen to be not assigned to the same type in the ontology. Notice that this mismatch is not due to accidental mistakes, but rather to the inherent multidimensional character of *ontolexical* systems, which may even require orthogonal principles of organization depending on the specific use contexts. These may actually impose different perspectives on the same conceptual system. Cognitive research has recently pointed out the fact that this context-dependency is an inherent feature of the human conceptual system (Barsalou, 2005). The phenomena of sense creation and semantic dynamics we saw above also point toward the same direction. Consistently, *ontolexical* resources can not be regarded as fixed repositories of semantic descriptions, but at most as core set of meanings that need to be customized and adapted to different domains, applications, texts, etc. NLP and learning methods can be used to achieve this sort of "textual attunement" of *ontolexical* resources, as a key condition to make them be better fitted for semantic processing tasks.

14.4 Conclusions

There is a natural connection between NLP and *ontolexical* resources. One of the main goal of the former is to understand the information and knowledge content that is encoded in natural language structures. The latter purport at representing knowledge systems that also happen to be expressed through natural

language expressions. The problem is that the naturalness of this link is often hindered by the way *ontolexical* knowledge is represented, organized and acquired. In fact, the effective usability of *ontolexical* systems for practical NLP tasks is not granted *per se*, and in some cases these resources have intrinsic limits that negatively impact on the way they can be used in processing tasks. This is the main reason why the two apparently independent moment of knowledge creation and knowledge use are inevitably interconnected within the "circle of life" of *ontolexical* systems. Indeed, NLP is part and a key protagonist of this circle. Better understanding how knowledge can automatically be carved out of texts can lead to *ontolexical* resources that are more "attuned" to the way knowledge is expressed with natural language. In turn, this promises to pave the way to better ways in which knowledge resources can be employed to boost NLP technology.