GUADALUPE AGUADO DE CEA
ELENA MONTIEL-PONSODA
Ontology Engineering Group, Universidad Politécnica de Madrid
Madrid, Spain
lupe@fi.upm.es
emontiel@delicias.dia.fi.upm.es

When Phraseology and Ontologies Meet

Introduction: Phraseology in Natural Language Processing Applications

In the field of Languages for Specific Purposes, phraseology plays a central role since its mastering is crucial for the different users that deal with specialized languages, from professionals of a domain to linguist experts (terminologists, translators, communication mediators, etc.). According to Aguado de Cea (2007), phraseology can be defined as: "the linguistic discipline that deals with the combination of words", (…), or "the set of phraseological units or phrasemes" of a certain specialized language. In any case, phrasemes or lexical combinations are not only necessary for an adequate specialized communication, but they are also fundamental in the understanding of a certain domain of knowledge, together with terms and the concepts underlying terms.

Therefore, from a cognitive point of view, phraseology can provide valuable insights in knowledge extraction. Recent research on Natural Language Processing (NLP) has also benefited from the application of phraseology to the acquisition of knowledge from corpora. If certain combinations of words allow a successful communication in a specific domain, they will also enable a successful acquisition of related knowledge. Research on NLP is not generally interested in specialized collocations, but it rather pays attention to those grammatical collocations that can be used to extract domain information in a semi-automatic or automatic way. In this paper, we will focus on those in which verbs are the central element in the phraseme, in that they also establish relations among several knowledge units. This kind of collocations have received several denominations in Terminology, such as "knowledge patterns" (Meyer, 2001), "conceptual relation patterns" (Condamines, 2002), or "linguistic markers" (Feliu and Cabré, 2002). In all these cases they have been used for speeding up the process of terminology extraction. For example, the combination of the lexical items "are types of" will be related to information about subcategories or subclasses of a certain concept. Consequently, if we introduce them in a knowledge extraction tool and search in a description text about *sensors*, the results obtained could show different categories or types of sensors, such as thermal, electromagnetic or chemical sensors.

However, in Terminology, not only are the concepts underlying terms important, but also how these terms are related to each other, what is represented by the so-called "conceptual relations". Feliu (2004: 27) has defined conceptual relations as elements "that link two or more specialized knowledge units in a particular subject field". In her work she analyzes the verbal phraseology of the Catalan language that expresses some of the most common conceptual relations, in order to apply them in terminology extraction (cf. Feliu and Cabré, 2002). In the same line, Marshman *et al*. (2002) focused on the identification of knowledge patterns in French for conveying some of the most usual relations in Terminology: hyperonymy (*est un* / *type de* / *forme de* [is a/type of/form of]), meronymy (*consiste en* / *partie de* / *comporte* [consist of/part of/includes]), and function (*utilise pour* / *permet* / *function* [is used for/allows/function]). In Computational Linguistics, Hearst (1992) introduced the concept of "Lexico-Syntactic Patterns" (LSP) to refer to some linguistic markers that enabled the identification of hyperonymy-hyponymy relations between concepts for accelerating the compilation of large lexicons. The set of LSPs that this researcher identified had the following characteristics: (1) they were directly extracted from texts, and (2) they had as main elements prepositional phrases, paralinguistic signs or conjunctions (but not verbs). E.g.: "…such as…", "…or other…", "…including…".

As we can see, the idea of automatically discovering related terms assuming a direct correspondence between lexical structures and conceptual or semantic relations (as Condamines (2002) also argued in her paper) appears recurrently in different disciplines. Unfortunately, this correspondence is not always true. Therefore, this poses some challenges to NLP researchers: the first one has to do with the efforts devoted to the discovering of reliable linguistic patterns that correspond to conceptual relations across domains and genres; and the second is related with language ambiguities (because of polysemy, for example), that have, as a consequence, the identification of various conceptual relations from the same linguistic pattern. In the following, we will explain how we attempt at contributing to the Ontology Engineering field with the use of Lexico-Syntactic Patterns, that we have redefined for our purposes as "linguistic schemas or constructs derived from recurrent expressions in natural language that consist of linguistic and paralinguistic elements that follow a certain syntactic order, and that permit to extract some conclusions about the meaning they express" (Aguado de Cea *et al.*, 2008). Then, we will focus on one of the most relevant ontological relation, the "subclass of" relation, and will present our proposals for facing the challenges that this relation raises because of the disparities between the lexical structures and the "subclass of" relation in an ontology. Finally, we will present the conclusions and some future lines of work.

Lexico-Syntactic Patterns in Ontology Engineering

Ontologies are one of the central research subjects in Artificial Intelligence and Knowledge Engineering, since they allow to represent knowledge for machines and to add semantics to the information in the Web. Moreover, ontologies represent a domain of knowledge by defining the concepts of that domain and the relations among them. So far, work on ontology development could be identified with terminology work. However, ontologies go some steps further, in the sense that definitions of concepts and relations among them are formalized, which

means that they are made understandable also to machines. And last but not least, the knowledge represented in an ontology captures the consensual knowledge of a community of domain experts. This has been summarized by Studer (1998: 161) in one of the most cited definitions that states that an ontology "(…) is a formal, explicit specification of a shared conceptualization". Broadly speaking, an ontology consists of four main components: concepts, attributes, relations and instances. Concepts identify types or classes of objects. Attributes refer to features or characteristics that define objects. Relations represent dependencies between concepts, or how concepts relate to each other. Instances are specific, real objects that belong to a certain class of objects.

As in Terminology, ontology development requires the discovery of the concepts of a specific domain, their properties, how they are related to each other, and the instances that belong to the identified concepts. Since this is a resource demanding and time consuming activity, many efforts have also gone to the automatic acquisition of the different ontology elements from free text. For this purpose, LSPs have been applied to extract ontology elements in order to speed up ontology development. Some researchers in this field (Snow 2004, Cimiano 2007 among others) have extended the original set of Hearst's patterns with additional ones that express hyponym relations, or new ones expressing relations such as meronymy, agency, cause, etc., for automatically enriching or populating ontologies from texts. Some patterns were similar to Hearst's ones, that is, not verb-centred, others had verbs as main elements. However, no research has been oriented to the identification and use of those LSPs equivalent to ontology relations with the aim of helping naive users in ontology development.

In this paper, we will also concentrate on verb-centred linguistic patterns, in line with Feliu and Cabré (2002) and Marshman *et al*. (2002).The main objective of this research is to create a repository of LSPs associated to the ontological relations they express. This repository will be stored in a system that will permit to identify when a sentence introduced by the user corresponds to an LSP, and in its turn, to an ontological relation, thus helping the user to construct an ontology. An overview of the system for automatically recognizing ontological relations is outlined in FIGURE 1. This system is currently being developed within the European Project NeOn[1].

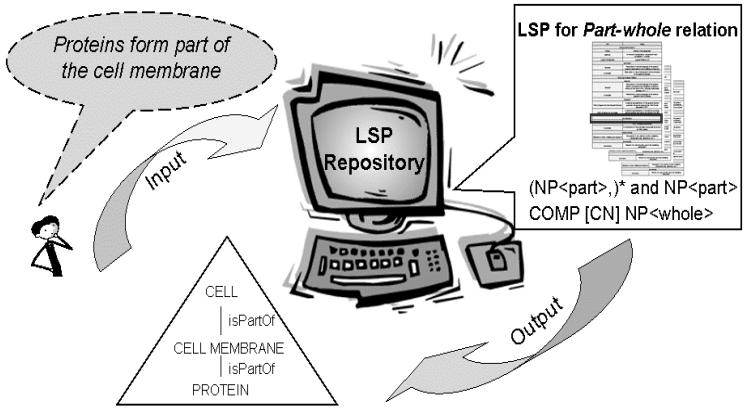

FIGURE 1. Overview of the System for LSP recognition.

Ontological Relations

Ontological relations, as they are understood in Ontology Engineering, can be divided into three main groups: taxonomic, meronymic and non-taxonomic relations. Taxonomic relations are hierarchical or inclusion relations, i.e., those that allow subordinate concepts to inherit the properties of the superodinate concept they belong to. They are also known as hyponymy or "subclass of" relations. Meronymic relations are the ones that hold between an object and its parts, also known as "part-whole" relations. The rest of ontological relations ("ad-hoc" relations of a specific domain) are considered non-hierarchical relations, for instance, "function" or "cause-result" relations.

In this paper, we want to approach one of the most significant relations both in Terminology and in Ontology Engineering, the "subclass of" relation. This is a fundamental relation whenever we want to organize the knowledge of a domain, and, in fact, it is the basic relation in the science of classification, i.e. taxonomy. Any concept related to another by the hyperonymy-hyponymy relation is said to inherit the properties of the hyperonym and has additional ones that specify it. This is also the basic idea behind the Aristotelian definition. Let us take as example the concepts "sensor" and "thermometer". We can say that a thermometer inherits the properties of the sensor in that it is "a device that measures a physical quantity and converts it to a signal", and specifies it by saying that the thermometer measures temperature.

Lexico-Syntactic Patterns for the "Subclass of" relation

In most ontology paradigms, the ontological relation "subclass of" indicates that "a class is a specialization of another class" (Gómez-Pérez *et al*., 2003: 49). In natural languages, this relation is present in definitions and realized linguistically, for instance, by means of the combination "…is a(n)…", as in *A thermometer is a sensor that* (…). In this sentence, the hyponym or subclass is the concept on the left-hand side of the verb, and the hypernym or superclass is on the right-hand side. By identifying the most used linguistic constructs that a language has for representing this kind of relation and including them in the LSP repository, we will allow users to

[1] http://www.neon-project.org/web-content/

express the “subclass of” relations present in their knowledge domain in full natural language (NL), and automatically model it in an ontology with the system previously outlined in FIGURE 1.

If we have a look at the following sentences in English, we will see that they establish a “subclass of” relation among the concepts in the sentence.

1. An orphan drug is a type of drug.

2. Membrane proteins are classified into two major categories: integral proteins and peripheral proteins.

We observed that the lexico-syntactic structures presented above appear recurrently across domains, and we decided to formalize them as LSPs. In TABLE 1 we include the LSPs representing these structures, and some additional ones we have identified for the English language. Some of the symbols and abbreviations used in the formalization of LSPs have been included below for the sake of readability.

LSP Identifier	LSP-Subclass of-English	
Formalization	1	[(NP<subclass>,)* and] NP<subclass> be [CN] NP<superclass>
	2	[(NP<subclass>,)* and] NP<subclass> (group in|into|as) | (fall into) | (belong to) CN NP<superclass>
	3	NP<superclass> CATV [CD] [CN] [PARA] (NP<subclass>,)* and NP <subclass>
	4	NP<superclass> be divide in|into | split | separate in|into [either] NP<subclass> or NP<subclass>
	5	There are CD CN NP<superclass> PARA [(NP<subclass>,)* and] NP<subclass>
Examples	1	*Odometry, speedometry and GPS are types of sensors.*
	2	*Thyroid medicines belong to the general group of hormone medicines.*
	3	*Membrane proteins are classified into two major categories, integral proteins and peripheral proteins.*
	4	*Sensors are divided into two groups: contact and non-contact sensors.*
	5	*There are two types of narcotic analgesics: the opiates and the opioids.*
Symbols and abbreviations: NP= Noun Phrase + semantic role between < > CATV = Verbs of Classification. This group includes verbs as: classify in/into, categorize in/into, sub-classify in/into, etc. CD = Cardinal Number CN= Class Name. This group includes: class of, group of, type of, member of, etc.		PARA = Paralinguistic symbols ( ) = groups two or more elements * = repetition [ ] = optional elements

TABLE 1. LSPs for the “subclass of” ontological relation

As has been mentioned, the linguistic structures presented above correspond to the ontological relation “subclass of”. However, from an ontological perspective it is recommendable to further specify this basic hierarchical relation by adding knowledge about “disjointness” and “exhaustiveness” of the hyponyms with regard to the hypernym. “Disjointness” is generally understood in ontological modelling as the property of two classes of not sharing subclasses or individuals. If we analyse sentence number 2, *Membrane proteins are classified into two major categories: integral proteins and peripheral proteins*, we should further determine whether “integral proteins” and “peripheral proteins” are two completely different groups and do not share any individuals, in other words, whether a protein belonging to the “integral proteins” group can also belong to the “peripheral proteins” class, or not. “Exhaustiveness” has to do with the property of a set of classes that belong to a superclass and include all individuals that belong to that superclass, without exclusion of any of them. Considering again sentence 2, we should also specify whether these two types of membrane proteins are the only types or groups in which membrane proteins can be divided, or if there is an additional type of membrane proteins.

The reason for this further specification of the "subclass of" relation has to do with ontology consistency. Taking into account that applications based on ontologies are able to reason with the information contained in ontologies, it is advisable to enrich the “subclass of” relations established in the ontology with information about disjointness and exhaustiveness to guarantee consistency checking and automatic evaluation of the elements contained in the ontology. To put it in other words, that sort of information helps checking if the ontology has been correctly instantiated. However, users who are not experts in ontology engineering are simply not aware of the fact that if they do not explicitly declare this kind of knowledge[2], some inconsistencies can result when reasoning with the ontology. For example, if subclasses are not declared to be disjoint, they may be considered to overlap, i.e., to share individuals. Both specifications of the “subclass of” relation may be implicit in the sentence, but they have to be made explicit in ontologies. In this sense, there have been some initiatives in order to automatically extract knowledge about disjointness from the natural language assertions expressed by users modelling ontologies. Among other features for automatically discovering disjoint subclasses, Völker *et al*. (2007), have intuitively identified two linguistic patterns in line with Hearst’s (1992) for the English language:

[2] The need of making explicit information about disjointness and exhaustiveness is necessary in ontology paradigms following Description Logics, because they rely on an open world assumption.

a) A pattern that contains the conjunction ...either…or… / …neither…nor… This conjunction indicates that the introduced concepts do not allow sharing instances, as they belong to only one of two groups.

b) A pattern based on enumerations. This pattern assumes that whenever users enumerate a set of concepts, these are pair wise disjoint.

There is no doubt that the function words “either” or “neither” undoubtedly refer to disjoint classes. Then, it can be stated that whenever the system comes across that conjunction, it will straightforward model knowledge units as “subclass of” relation and “disjoint classes” in the ontology. For this purpose, we have created new patterns that directly identify both relations. See TABLE 2 below.

LSP Identifier	LSP-Subclass of + Disjoint classes-English	
Formalization	1	NP<superclass> be | CATV either NP<subclass> or NP<subclass>
	2	NP<superclass> be | CATV neither NP<subclass> nor NP<subclass>
Examples	1	*Animals are either vertebrates or invertebrates.*

TABLE 2. LSPs for the “subclass of” plus “disjoint classes” ontological relations

According to the enumeration assumption, we could contend that the identified LSPs 3, 4 and 5 for the “subclass of” relation (see TABLE 1) could also be regarded as patterns indicating additionally that subclasses are pair wise disjoint. In order to validate this statement we analysed the use of some verbs of classification in the British National Corpus[3] (BNC). The verb forms searched in the BNC were: “classified into”, “divided into”, “split into” and “separated into”. Although this is an ongoing research, at this stage we observed that out of the sentences retrieved with the meaning of "subclass of"[4] (50 sentences analysed for each verb form), between 80% and 90% clearly expressed disjoint subclasses according to domain knowledge. This statement was further supported by the fact that sentences often included a cardinal number as well as adjectives like “distinct” or “separate” accompanying class names, e.g.:

- “Covalent crystals are classified into two distinct groups…”
- “Government institutions are divided into two separate scientific communities…”

The results of this initial analysis confirm that when users provide an enumeration of subclasses, these are normally pair wise disjoint. Nevertheless, in order to cope with the remaining 10% to 20% of cases in which information about disjointness was not so clear, a system to interact with the user has been planned. This will enable users to become aware of the relevance of making explicit that sort of information when modelling ontologies.

Thus, our system formulates some questions in order to find out whether subclasses are disjoint. Taking as example sentence 2 *Membrane proteins are classified into two major categories: integral proteins and peripheral proteins*, the question launched by the system would be:

- Can a certain membrane protein belong to the category of integral proteins and peripheral proteins at the same time?

In this case, the answer should be “no”, and the system would further model those subclasses as disjoint classes. If the answer is yes, it would just model them as subclasses of membrane proteins.

Regarding exhaustiveness, whenever there is a cardinal number, as in example 2, we can be nearly sure that the enumeration is complete. However, if this cardinal is not present in the sentence, one option to find out whether the listed classes are exhaustive could be by answering the following question posed by the system:

- Are there any other types of membrane proteins?

According to our knowledge of the domain, the question should be “no”, in order to be completely sure that the right modelling decision is to further model those classes as exhaustive classes.

Conclusions

In this paper we have tried to show how grammatical collocations can provide interesting hints about the conceptual relations underlying them. In fact, these combinations of lexical items have been used in fields like Terminology and Ontology Engineering with the purpose of automatically extracting data for accelerating respectively terminology tasks and ontology development. However, with this new approach to what we have called Lexico-Syntactic Patterns (LSPs), we aim at applying lexical combinations to help users who are not experts in ontology engineering to develop ontologies by formulating in NL what they want to model. In order to support the system for an automatic identification of LSPs corresponding to ontological relations, we have developed a repository of LSPs associated to ontological relations that make up the core of the system. In this paper, we have focused on those LSPs expressing the "subclass of" ontological relation, and have pointed out the semantic differences between the lexical combinations expressing this relation and the “equivalent” ontological relation, since the latter needs to be further specified with information about disjointness and exhaustiveness. As

[3] http://www.natcorp.ox.ac.uk/

[4] Note that some of these verbs (e.g. divide in/into) are ambiguous and could also indicate a “part-whole” relation. For more information, see (Montiel-Ponsoda, *et al*., 2008, and Aguado de Cea *et al*., 2008).

explained in the paper, some lexical elements may help to directly identify these specifications. Notwithstanding, with the aim of validating and making explicit these properties of the "subclass of" relation, user interaction with the system has been devised. Future work will be centred on enriching this initial repository of LSPs, putting special emphasis on the discovery of disparities between lexical and ontological relations.

Acknowledgements. This research has been supported by the European project *NeOn* (FP6-027595), and the National project *GeoBuddies* (TSI2007-65677C02).

References

Aguado de Cea, G., Gómez-Pérez, A., Montiel-Ponsoda, E., Suárez-Figueroa, M.C. (2008). "Natural Language-based Approach for Helping in the Reuse of Ontology Design Patterns". To appear in *EKAW 2008*, Catania, Italy.

Aguado de Cea, G. (2007). "A Multiperspective Approach to Specialized Phraseology: Internet as a Reference Corpus for Phraseology". In S. Posteguillo, M.J. Esteve and M.L. Gea-Valor (eds.). *The Texture of Internet*: *Netlinguistics in progress*. Newcastle: Cambridge Scholars Publishing.

Cimiano, P. and Wenderoth, J. (2007). "Automatic Acquisition of Ranked Qualia Structures from the Web". In *Proc. of the Annual Meeting of the Association for Computational Linguistics,* 888--895.

Condamines, A. (2002). Corpus analysis and conceptual relation patterns. *Terminology*. Vol. 8. (1), 141-162.

Feliu, J. and M.T. Cabré. (2002). "Conceptual relations in specialized texts: new typology and an extraction system proposal". In *Proc. of TKE 2002*. Nancy, 45-49.

Feliu, J. (2004). *Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica*. PhD Thesis. Institut Universitari de Lingüística Aplicada.

Gómez-Pérez, A., Fernández-López, M. Corcho, Ó. (2003). *Ontological Engineering*. Springer, New York.

Hearst, M. A. (1992). "Automatic Acquisition of Hyponyms from Large Text Corpora". In *Proc. of the14th International Conference of Computational Linguistics*, 539-545.

Meyer, I. (2001). "Extracting knowledge-rich contexts for terminography. A conceptual and methodological framework". In C. Bourigault, (ed.), *Recent Advances in Computational Terminology*, 279-303. Benjamins.

Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., Suárez-Figueroa, M.C. (2008). "Helping Naive Users to Reuse Ontology Design Patterns". In *Proc.of the KRRSW*, co-located at the ESWC2008, in Tenerife, Spain.

Snow, R., Jurafsky, D., Ng, A. Y. (2004). "Learning syntactic patterns for automatic hypernym discovery". In *Advances in Neural Information Processing Systems* 17.

Studer, R., Benjamins, R., Fensel, D. (1998). "Knowledge engineering: principles and methods". In *Data & Knowledge Engineering 25 (1-2)*, 161-198.

Völker, J., Vrandecic, D., Sure, Y., Hotho, A. (2007) "Learning Disjointness". In Enrico Franconi and Michael Kifer and Wolfgang May,(eds.), *Proc.of the ESWC2007*, Springer-Verlag.