Using Natural Language Patterns for the Development of Ontologies

Elena Montiel-Ponsoda
Guadalupe Aguado de Cea
Universidad Politécnica de Madrid.

Abstract

The combination of certain linguistic units that recurrently appear in text genres has attracted the attention of many researchers in several domains, as they can provide valuable information about different types of relations. In this paper, the focus will be on some of these combinatory units, referred to as Lexico-Syntactic Patterns (LSPs) that provide information about conceptual relations. The aim of this research is to detect recurrent patterns that express some of the most common conceptual relations present in ontologies. The purpose of this paper is to present the different strategies we have followed to identify LSPs which correspond to some of the main ontological relations, as well as an excerpt of the repository of LSPs that is currently being built.

Key words: conceptual relations, Lexico-Syntactic Patterns, ontological relations.

Resumen

Las combinaciones de ciertas unidades lingüísticas que aparecen de forma recurrente en textos de diferentes géneros han atraído la atención de muchos investigadores. Este artículo se centra en algunas de estas combinaciones, a las que hemos denominado Patrones Léxico-Sintácticos (LSPs en sus siglas en inglés), y que proporcionan información muy valiosa sobre relaciones conceptuales. El objetivo de esta investigación es identificar aquellos patrones que reflejan las relaciones conceptuales que encontramos en las ontologías. En este artículo presentamos las principales estrategias que hemos adoptado para el descubrimiento de LSPs

que expresan relaciones ontológicas, así como un extracto del repositorio de LSPs que estamos desarrollando.

Palabras clave: relaciones conceptuales, Patrones Léxico-Sintácticos, relaciones ontológicas.

Introduction

When describing recurrent combinations of certain linguistic units, researchers have taken into account the morpho-syntactic, semantic and pragmatic features that collocations present. According to Aguado (2007: 182), depending on the combination of criteria adopted in collocational studies, the approaches comply mainly with syntactic criteria, lexico-syntactic criteria, semantic criteria and pragmatic criteria, though other criteria such as statistical and conceptual (Hearst, 1992, Meyer & Mackintosh 1996, Feliu & Cabré 2002) have also been applied.

In Terminology, conceptual relations play a decisive role, since they illustrate "the network of concepts underlying the terms of a domain" (Meyer, 2001: 280). Concepts and conceptual relations are the basic research objects in terminology (Cabré *et al*., 1996), together with their linguistic realizations, i.e. the terminological units. As Meyer (2001: 280) reports, there are two types of conceptual characteristics: (1) attributes (e.g. colour, height, weight) that hold for a certain concept without involving other concepts, and (2) relations, which link concepts and help to describe domain knowledge (e.g. hyponymy, meronymy, causality).

With the aim of describing a domain of knowledge we need to find out, first, the relevant concepts of the domain and their description, and second, how they are related to each other. In this paper we will deal with conceptual relations in that they are centred on how world objects are related to each other and on the lexical realizations that convey a certain relation between them.

Identifying Conceptual Relations by means of Linguistic Markers

Conceptual relations are defined by Feliu (2004: 27) as elements "that link two or more specialized knowledge units in a particular subject field", and they are formally represented as:

R (a b, n)

where "R" represents the relation, "a" and "b" are knowledge units, and "n" foresees the case when a relation links more than the two elements "a" and "b". In her work, she analyzes verb-oriented conceptual relations, i.e., those relations in which verbs are the ones that convey a specific relational meaning. The objective of her research was to detect those "linguistic patterns" that expressed conceptual relations and to apply them to terminology extraction. A catalogue of linguistic patterns conveying the relations of similarity, inclusion, sequentiality, causality, instrument, meronymy and association for the Catalan language is included in Feliu and Cabré (2002). With a similar objective, Meyer (2001: 290) identified "knowledge patterns" with the aim of extracting terminology in a semi-automatic way.

In Marshman *et al*. (2002) knowledge patterns are defined as "words, word combinations or paralinguistic features of texts which frequently indicate conceptual relations", and are divided in:

- Lexical knowledge patterns, which consist of words or groups of words.
- Grammatical knowledge patterns, which involve combinations of grammatical categories.
- Paralinguistic knowledge patterns, which are neither lexical nor grammatical, but include elements of text such as punctuation or parentheses.

Marshman *et al*. (2002) focused on the identification of "lexical knowledge patterns" used in French for conveying three types of conceptual relations: 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) also identified some linguistic markers with the goal of extracting information that would help in building up large lexicons for natural language processing. Hearst mainly focused on the automatic acquisition of hyponymy relations from texts by means of what she called "Lexico-Syntactic Patterns" (LSP henceforth). Hearst's LSPs are said to "occur frequently and in many text genres, almost always indicate a relation of interest, and be recognized with little or no pre-encoded knowledge". 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). Examples of Hearst's patterns are shown in Table 1.

NP^1 *such as* $\{NP_1, NP_2 \ldots$ *(and | or)* NP_n
NP {,NP} {,} or other NP*
NP {,NP} {,} or other NP*
NP {,} including {NP,} { or | and } NP*
NP {,} especially {NP,} { or | and } NP*

Table 1. Hearst's patterns

Since then, there have been many authors that have applied Hearst's LSPs for the automatic discovery of lexical items. In the next section, we will deal with the use of LSPs in the Ontology Engineering field.

Lexico-Syntactic Patterns in Knowledge 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" (based on Gruber (1993)).

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. Consider an ontology of animals, where "mouse" would be a type or subclass of "animal", i.e., a concept in the ontology; "size", "weight" and "colour" of "mouse" would be the attributes; "mouse" could be related to "cheese" by means of the relation "eats"; and a certain mouse called "Mickey" could be an instance of the concept "mouse".

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 time and resource consuming activity, many efforts have gone to the automatic acquisition of the different ontology elements from texts. For this

purpose, LSPs have been applied to extract ontology elements in order to speed up ontology development. Some researchers, Snow (2004), or 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. Some patterns were similar to Hearst's ones, that is, not verb-centred, others had verbs as main elements. In any case, no research has been oriented to the identification and use of those LSPs that are 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).

For this purpose, we have also adopted the name of LSPs, but we have redefined them 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" (Montiel-Ponsoda *et al.*, 2008). 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[7].

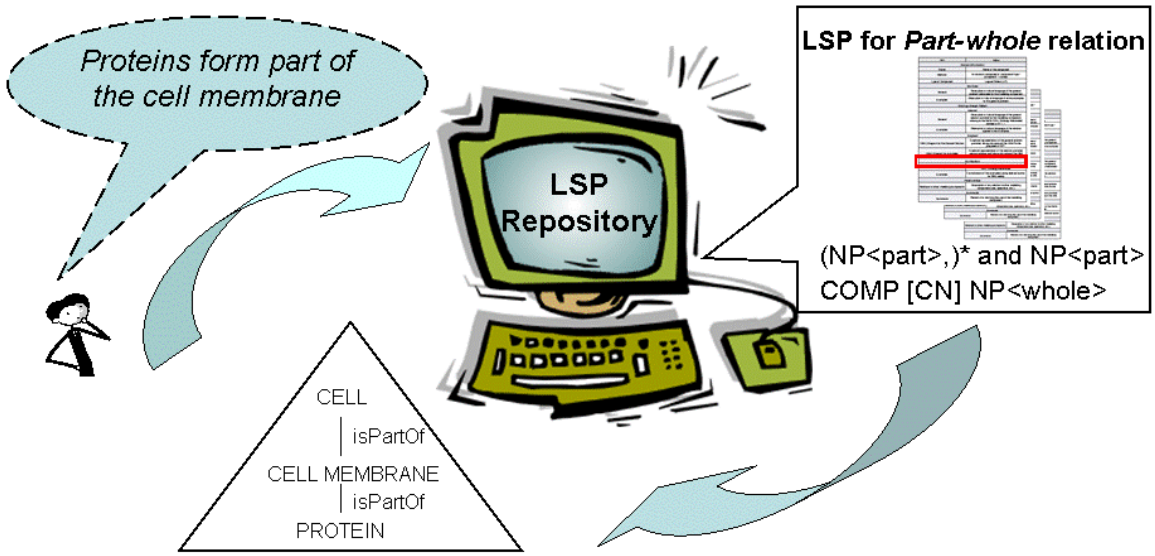

Figure 1. Overview of the System for LSP recognition.

From an ontology engineering viewpoint, ontological relations can be divided into two main groups: taxonomic and non-taxonomic relations. In ontologies, the most important ones are the taxonomic or hierarchical relations, because they allow subordinate concepts to inherit the properties of the superodinate concept they belong to. They are also known as hyponymy or "subclass of" relations. The rest of ontological relations ("ad-hoc" relations of a specific domain) are considered non-taxonomic relations. For instance, the relation expressed by the verb "to eat" in "mouse eats cheese", is a non-taxonomic relation. There are also other types of relations, meronymic or "part-whole" relations, that *strictu sensu* are not taxonomic, and they express the relations held between an object and its parts. In the next sections, we give an overview of the main strategies followed for the extraction of LSPs, as well as some examples of LSPs for the "subclass of" relation, and the "part-whole" relation.

Strategies in the Discovery of LSPs

At this stage, the methodology applied for extracting natural language expressions equivalent to ontological relations, and transformed afterwards in LSPs, conformed to the following strategies:

a) To select available verb-oriented patterns in literature and adapt them to our notation schema, following symbols and abbreviations based on a well known notation form in Computer Science, the Backus-Naur Form[2] (see Table 2).

For instance, one of Cimiano's patterns (2007) that expresses the "subclass of" relation, e.g., "NP_{QT} is a kind of NP_F" would be transformed into "NP<subclass> be [CN] NP<superclass>", according to our notation (see also Table 3).

b) To identify ontology related concepts, and search in the Web for common verbal constructs that link them. As input resources we used WordNet[3]and the AGROVOC[4] thesaurus, that include concepts related among them.

For example, we took the words "protein" and "cell" from WordNet conscious of the "part-whole" relation that holds between them, and introduced both terms in the Web. Results were sentences like: "A typical human cell contains millions of these proteins (…)", corresponding to our LSP "NP<whole> have | contain (NP<part>,)* and NP<part>" (see Table 4).

SYMBOLS & ABBREVIATIONS	DESCRIPTION
AP<...>	Adjectival Phrase. It is defined as a phrase whose head is an adjective accompanied optionally by adverbs or other complements as prepositional phrases. AP is followed by the semantic role played by the concept it represents in the conceptual relation in question, such as e.g., *property*.
CATV	Verbs of Classification. Set of verbs of classification plus the preposition that normally follows them. Some of the most representative verbs in this group are: *classify in/into, categorize in/into, sub-classify in/into*.
CD	Cardinal Number.
CN	Class Name. Generic names for semantic roles usually accompanied by preposition, such as *class, group, type, member, subclass, category, part, set,* etc.
COMP	Verbs of Composition. Set of verbs meaning that something is made up of different parts. Some of the most representative ones are: *consist of, compose of, make up of, form of/by, constitute of/by*.
NP<...>	Noun Phrase. It is defined as a phrase whose head is a noun or a pronoun, optionally accompanied by a set of modifiers, and that functions as the subject or object of a verb. NP is followed by the semantic role played by the concept it represents in the conceptual relation in question, e.g., *class, subclass, or property*.
PARA	Paralinguistic symbols like *colon*.
()	Parentheses group two or more elements.
*	Asterisk indicates repetition.
[]	Elements in brackets are meant to be optional, which means that they can be present either at that stage or not, and by default of appearance, the pattern remains unmodified.
¬	Elements preceded by this symbol should not appear in the pattern.

Table 2. Restricted symbols and abbreviations in LSPs

c) To search in domain descriptive and encyclopaedic documents for verbal constructs that link concepts according to ontological relations (For this goal, we used descriptive documents from the Web and documentation used in the development of ontologies in some European Projects such as Esperonto[5] and OntoGrid[6]).

As a result of applying these strategies we obtained a set of sentences in which concepts were related by means of different verbal constructs. We could state that some sentences corresponding to the same ontological relations followed a similar schema, despite concepts

coming from different knowledge domains. For illustrating this process, let us consider two sentences in English from different knowledge domains expressing a “subclass of” relation:

1. Animals are divided into two major categories: vertebrates and invertebrates.
2. Medications are generally classified into two groups: over-the-counter (OTC) medication and prescription only medicines (POM).

The verbs “divide into” and “classify into” indicate a subclass of relation, in which the so-called “superclass” is at the left-hand side of the verb, and the “subclasses”, at the right-hand side. In fact, there is a group of sentences constructed in a similar way from which we could draw an LSP embracing all of them:

NP<superclass> CATV [CD] [CN] [PARA] (NP<subclass>,)* and NP <subclass>

However, the verb “divide” in the first sentence could indicate a “part-whole” relation, in certain contexts, as in “The cerebrum is divided into two major parts: the right cerebral hemisphere and left cerebral hemisphere”. Therefore, it has to be excluded from the above identified LSP, and considered a special case of ambiguous LSP that can correspond to two different ontological relations. See Table 5 for more examples.

Lexico-Syntactic Patterns for the “Subclass of” and the “Part-whole” relations

In the following tables we show an excerpt of the LSP repository we are currently building in order to collect the different ways of expressing ontological relations in a language. For each of the identified LSP, we have added an example in English. Although for the moment being, the set of LSPs discovered and associated to each ontological relation is not exhaustive, it aims at being representative of the most typical ways in which a language can express this

relations. In Table 3, we find the set of LSPs for the “subclass of” ontological relation in the English language, as the LSP Identifier indicates.

LSP Identifier	LSP-*subclass of*-EN	
Formalization	1	NP<subclass> be [CN] NP<superclass>
	2	[(NP<subclass>,)* and] NP<subclass> be [CN] NP<superclass>
	3	[(NP<subclass>,)* and] NP<subclass> (group in|into|as) | (fall into) | (belong to) CN NP<superclass>
	4	NP<superclass> CATV [CD] [CN] [PARA] (NP<subclass>,)*and NP <subclass>
	5	There are CD CN NP<superclass> PARA [(NP<subclass>,)* and] NP<subclass>
Examples	1	*An orphan drug is a type of drug.*
	2	*Odometry, speedometry and GPS are types of sensors.*
	3	*Thyroid medicines belong to the general group of hormone medicines.*
	4	*Membrane proteins are classified into two major categories, integral proteins and peripheral proteins.*
	5	*There are two types of narcotic analgesics: the opiates and the opioids.*

Table 3. LSPs for the “subclass of” ontological relation

LSP Identifier	LSP-*part-whole*-EN	
Formalization	1	(NP<part>,)* and NP<part> COMP [CN] NP<whole>
	2	NP<whole> be COMP [CN] (NP<part>,)* and NP<part>
	3	NP<whole> have | contain (NP<part>,)* and NP<part>
Examples	1	*Proteins form part of the cell membrane.*
	2	*A state machine workflow is made up of a set of states, transitions and actions.*
	3	*Cars have tires*

Table 4. LSPs for the ”part-whole” ontological relation

However, identifying LSPs is not a trivial process, since certain polysemous verbs can correspond to several ontological relations, which are clear for humans, but not for machines, as mentioned in the previous section. In the following table we have included some

ambiguous LSPs, i.e., LSPs that can correspond to two different types of relations, namely, “subclass of” relation and “part-whole” relation. Machines could interact with humans in order to disambiguate the verb sense, and find out which ontological relation the user is referring to. Research is currently being conducted in this sense.

LSP Identifier	LSP-*subclass of-part-whole*-EN	
Formalization	1	NP<class> include | comprise [(NP<class >,)* and] NP<class>
	2	NP<class> be divided | split | separate in|into | [CN] [(NP<class >,)* and] NP<class>
Examples	1	*Arthropods include insects, crustaceans, spiders, scorpions, and centipedes. (subclass of)* *Reproductive structures in female insects include ovaries, bursa copulatrix and uterus. (part-whole)*
	2	*Marine mammals are divided into three orders: Carnivora, Sirenia and Cetacea. (subclass of)* *The cerebrumis divided into two major parts: the right cerebral hemisphere and left cerebral hemisphere. (part-whole)*

Table 5. LSPs for the “subclass of” and “part-whole” ontological relations

Conclusions

Linguistic patterns in Terminology and Knowledge Engineering have proven to be highly beneficial for extracting valuable information for speeding up terminology and ontology work. Our proposal of identifying Lexico-Syntactic Patterns (LSPs) that correspond to ontological relations can help users in the development of ontologies by using a system that permits an automatic detection of the ontological relation expressed in the sentence introduced by the user. The core of this research is the repository of LSPs that will enable to automatically recognize ontological relations. When developing systems for the automatic recognition of patterns from expressions in natural language, we have to cope with some features of natural language that require special effort, such as, language polysemy, as shown in this paper. At

present, the LSP repository is being extended to cover all these ontological relations. We plan to enhance the repository with LSPs for Spanish and German.

Acknowledgements

The research described in this paper is supported by the project *Lifecycle support for networked ontologies (NeOn)* (FP6-027595). In addition, it is partially co-funded by an I+D grant from the *Universidad Politécnica de Madrid*. We also would like to thank Mari Carmen Suárez-Figueroa for their help with some ontological aspects.

References

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.

Cabré, M. T., J. Morel, and C. Tebé. (1996). "Las relaciones conceptuales de tipo causal: un caso práctico". Proceedings of the V Simposio Iberoamericano de Terminología: Terminología, ciencia y teconología. México: Unión Latina, 82-94.

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

Feliu, J. and M.T. Cabré. (2002). "Conceptual relations in specialized texts: new typology and an extraction system proposal". TKE2002. Nancy, 45-49.

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

Hearst, M. A. (1992). "Automatic Acquisition of Hyponyms from Large Text Corpora". 14th International Conference on Computational Linguistics, 539-545.

Meyer, I. and K. Mackintosh. (1996). "Refining the terminographer's concept-analysis methods: How can phraseology help?" *Terminology* Vol 3(1), 1-26.

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". To appear in *Proceedings of the 1st International Workshop on Knowledge Reuse and Reengineering over the Semantic Web*, at the European Semantic Web Conference (ESWC08), 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.

Gruber, T.R. (1993). "Toward principles for the design of ontologies used for knowledge Sharing". In N. Guarino y R. Poli, (eds.), *International Workshop on Formal Ontology*, Padova, Italia.

Notes

[1] NP: Noun Phrase

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

[3] http://cui.unige.ch/db-research/Enseignement/analyseinfo/AboutBNF.html

[4] http://wordnet.princeton.edu/

[5] http://www.fao.org/aims/ag_intro.htm

[6] http://www.esperonto.net/semanticportal/jsp/frames.jsp

[7] http://www.ontogrid.net/ontogrid/publications.jsp