In Search for More Knowledge: Regular Polysemy and Knowledge Acquisition

Wim Peters
University of Sheffield
Regent Court, Portobello Street
Sheffield S1 4DP
United Kingdom
w.peters@dcs.shef.ac.uk

Abstract

This paper describes the process of the extraction of implicit knowledge from WordNet and EuroWordNet. This knowledge is an extension of the explicit knowledge structures already provided by the wordnets in the form of synsets and semantic relations, and is contained both within (Euro)WordNet's hierarchical structure and the glosses that are associated with each WordNet synset. The extended knowledge comes in the form of frame structures containing regular polysemic patterns and automatically extracted relations that link the participating concepts in these patterns.

1 Introduction

WordNet (Fellbaum, 1998) and EuroWordNet (Vossen, 1998) are popular resources for a number of NLP tasks such as semantic tagging (Volk et al, 2002) and information retrieval (Gonzalo et al., 1998). Although the thesauri contain a wealth of information that can be put to use for these purposes, only part of their lexical knowledge is available in explicit form, organised in synsets and a number of semantic relations between these synsets such as synonymy, hypernymy and thematic relations. NLP applications can only exploit this type of lexical knowledge, although it is just a part of what these thesauri can offer. Much additional information is hidden away in (Euro)WordNet's ontological structure, and within the glosses that are associated with the synsets.

This paper describes an attempt to tease out at least part of this implicitly available lexical knowledge.

The area we concentrate on is that of figurative language use, in particular cases of regular polysemy (Apresjan, 1973). This type of figurative language use exploits semantic regularities in language that, when captured, offer valuable additional semantic information for the concepts involved, and complement the psycholinguistically oriented knowledge explicitly encoded in WordNet.

Viewed traditionally, regular polysemy (henceforth RP) is a metonymic phenomenon: a non-literal figure of speech in which the name of one thing is substituted for that of another related to it. In its basic form, it establishes a semantic relation between two concepts that are associated with the same word. RP is regular in that it captures conventionalized and therefore recurrent processes of sense extension. In general, RP is lexicalized, i.e. they are explicitly listed in dictionaries and as such more independent of a pragmatic situation (as opposed to irregular sense extensions, which can be created on the fly and are determined by pragmatic constraints in discourse). For example, the *White House* is, on the one hand, an institution and, on the other, a building. The semantic relation between the two senses is 'is housed in'. This relation is also applicable between conceptually equivalent lexicalized senses of other words, such as *school*, *college*, *academy* and *hospital*. Therefore, the constellation of lexicalized senses and conventionalized relations is called regular polysemy. It is quite possible that *school* might have, in some contexts, acquired the sense of 'place of torture', which falls outside the scope of RP, because this sense is pragmatically restricted, not lexicalized and not more widely applicable to other words. Similarly, given the present world situation, *White House* has possibly acquired other senses as well, which are not taken into account in this study because of they have not been conventionalized yet.

Examples of regular polysemy have until now been the product of linguistic introspection and manual lookup in dictionaries and texts. Automatic approaches such as Buitelaar (1998) only concentrate on patterns associated with high level concepts within the WordNet hierarchy. In general, the availability of electronic semantic resources such as WordNet makes it possible to extract and investigate regularities between sense distinctions in a data-driven way. These regularities form the core data set for the derivation of extended knowledge fragments that complement the (Euro)WordNet knowledge structure.

2 Automatic Selection of Regular Polysemy

The work consisted of three phases. First, an automatic selection process identified candidates for instantiations of regular polysemy (For a detailed description see Peters, 2002 and Peters and Wilks, 2003) in WordNet on the basis of systematic sense distributions of nouns. These systematic distributions can be characterized by a pair of hypernyms taken from the WordNet hierarchies that subsume the sense combinations of the words involved. For instance, in two of its senses 'law' falls under the pattern *profession* (an occupation requiring special education) and *discipline* (a branch of knowledge). This pattern is also displayed by four other words in WordNet, namely 'architecture', 'literature', 'politics' and 'theology'. Figure 1 illustrates this case.

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 245–250. © Masaryk University, 2005

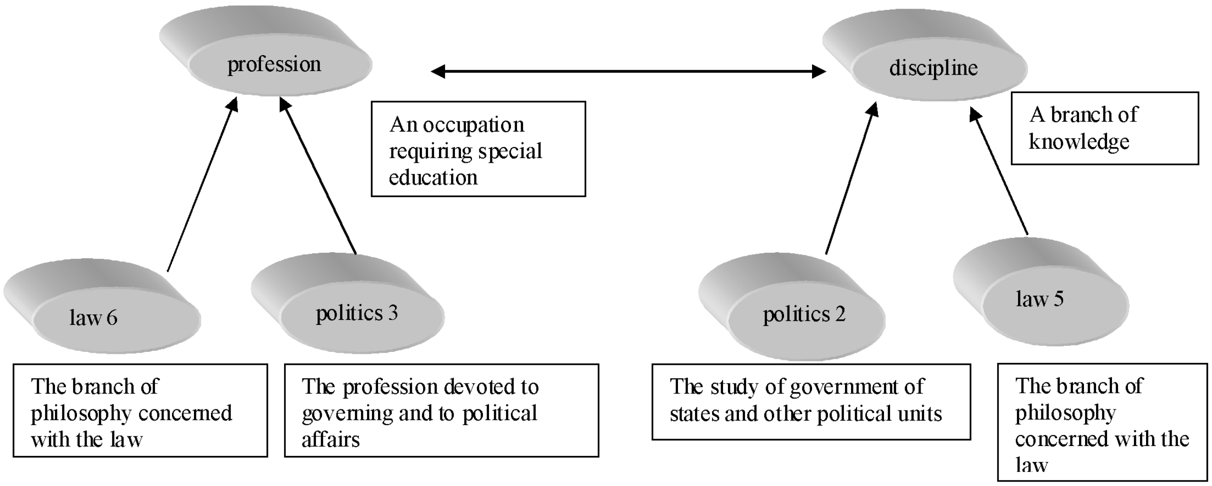
profession
discipline
An occupation requiring special education
A branch of knowledge
law 6
politics 3
politics 2
law 5
The branch of philosophy concerned with the law
The profession devoted to governing and to political affairs
The study of government of states and other political units
The branch of philosophy concerned with the law

Figure 1: The Regular Polysemic Pattern *Profession – Discipline*

3 Extraction of Semantic Relations

There are several ways in which the determination of the semantic relations between the related word senses can be approached. First, one can stop here and use the unlabelled associations in a knowledge base by employing a semantic relation such as ‘is related to’ between the senses. The semantic characterization of the relation will then only imply that they are related in some unspecified way. Second, these relations can be determined by means of human introspection. The examination of the pair and the participating word senses will provide a human assessor with enough information to postulate a relationship. However, this is a costly and time consuming activity that is worthy of a project on its own. An example of this type of project is FrameNet (Fillmore and Atkins, 1992), which involves as one of its phases the determination of the structure of the frames involved, and the nature of the slots and fillers.

A third option is to automate this process by exploiting the semantic information available in the WordNet glosses for the process of extracting explicit semantic relations between the word senses involved in the regular polysemic pattern. This option was explored in the second stage, in which the relations that exist between the word senses that participate in patterns were acquired in an automatic fashion. This additional information is obtained by analyzing the glosses that are associated with the synsets of the word senses involved and their hypernyms. After part of speech tagging and lemmatization all synonyms and hypernym synset members of the participating words are grouped together into two bags of words. These are then mapped onto the glosses that bare associated with all synsets in the hypernymic paths. If a verb occurs between pairs of words from each bag this is taken as the semantic relation that holds between the word senses. Figure 2 illustrates the process.

We exemplify the results by means of the pattern *profession* (an occupation requiring special education) and *discipline* (a branch of knowledge) described above. Sense 6 of *law* has the gloss ‘the learned profession that is mastered by graduate study in a law school and that is responsible for the judicial system; "he studied law at Yale"’

Bag synset 1 contains *profession*, bag synset2 *study*. In between is the verb ‘is mastered by’ which yields the relation for this regular polysemic pattern. By adding thematic roles to the concepts involved one can state that *discipline* is either the subject or instrument associated with ‘master’ and *profession* is the object. Overall, relations have been extracted for around 5000 candidate regular polysemic patterns. A few more examples are listed in Table 1.

The explicit relations between word senses that can be gleaned from information implicit in glosses enriches the existing knowledge structures of WordNet in a horizontal way, i.e. they add ontological information that is not taxonomic in nature, thereby expanding its coverage as a knowledge base.

Also, these sets form the start of the explicit encoding of metonymic potential of words that do not yet participate in the patterns. For instance, *August plum*, *avocado* and *bergamot* are hyponyms of *fruit tree* (amongst about 300 more), but do not have an *edible fruit* sense. This sense can be postulated on the basis of the attested RP pattern.

4 Creation of Extended Knowledge Fragments

In the third phase, increasingly larger knowledge structures are built up on the basis of the sense pairs involved in the attested patterns. The frame model (Minsky, 1975; Fillmore, 1977), elaborated in psychology and linguistics in the last two decades, provides us with a well established formalism for the illustration and representation of this systematic knowledge. Frames are conceptual wholes that can be found under various denominations in the literature, such as scenes, scenarios, domains, and Idealized Conceptual Models (Lakoff, 1987). The frame structure is defined as the relation that exists between elements of a frame or between the frame as a whole and its elements. The relation triples

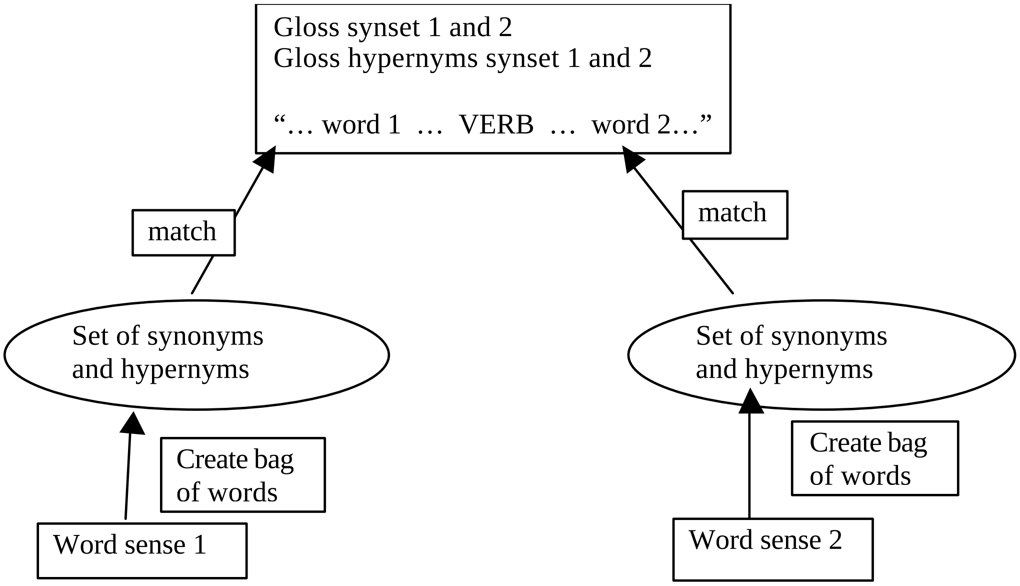
Gloss synset 1 and 2
Gloss hypernyms synset 1 and 2
"… word 1 … VERB … word 2…"
match
match
Set of synonyms and hypernyms
Set of synonyms and hypernyms
Create bag of words
Create bag of words
Word sense 1
Word sense 2

Figure 2: Extracting relations between the word senses

extracted in the second stage (e.g. **person-speak-language**) form the basic building blocks of the frames.

Extension of these rudimentary frames takes place in two ways. First, the concepts with which a hypernym from a particular regular polysemy pattern co-occurs can be regarded as additional slots in a topical frame that characterizes that hypernym. For instance, the pattern 'music-dance' covers words such as *tango* and *sonata*. Music in its turn co-occurs with a number of other concepts within the hypernym pairs that capture the regular polysemic patterns. A number of examples are listed in Table 2.

These concepts and the relations that have been extracted between these hypernyms form a further extension of the MUSIC frame. For music, the following relations with other hypernyms have been extracted: **person-make/accomplish-music** and **music-accompany-activity**.

Secondly, a further extension takes the semantic context of EuroWordNet into account. From the superset of all concepts and relations that are linked to MUSIC in all eight language specific wordnets the MUSIC frame is extended with this knowledge. The resulting frame structure is illustrated by Figure 3.

5 Conclusion and looking ahead

The applied methodology demonstrates that it is feasible to enrich the explicit structure of (Euro)WordNet with information that is implicitly available in the resource itself. This information augments the semantic coverage of (Euro)WordNet, and incrementally extends the knowledge base.

It is commonly agreed that a substantial knowledge base is one of the key necessities for natural language understanding in order to bootstrap the interpretation process. This knowledge base should contain enough information to allow the application of inferencing methods for the processing of coreference, anaphora and bridging expressions. For instance, for a text segment such as:

"The composer finished his sonata. Music had always been his first love.",

the extended knowledge fragment for *music*, illustrated in figure 3, enables the detection of linking relations between the nouns (*composer*, *sonata*, *music*). This is achieved by means of the hypernymic relation between *sonata* and *music*, another hypernymic relation between *composer* and *person*, and the 'make' relation between *person* and *music*.

No available resource is complete in its coverage of semantic knowledge. WordNet therefore suffers the same fate, and the fact that the characterization is wholly dependent on WordNet's sense distinctions and taxonomic structure of course influences the results. In order to balance out this bias towards WordNet and its sense distinctions, the applied techniques for extraction can be applied to any resource with taxonomic information, and it can be expected that regular polysemic patterns from different resources might well complement the WordNet derived data is some ways.

In general, the ideal knowledge base should contain as much lexical semantic information as possible in order to identify standardized, i.e. lexicalized, uses of words and their relations on the one hand, and enable the interpretation of non-standard usage on the other. An appropriate model must contain both particular knowledge about some non-standard interpretation, and reasoning to make the non-standard interpretation fit the current context. The described methodology shows the feasibility of capturing at least part of the knowledge necessary for forming the appropriate interpretation of a word in a text, in the form of an attested or potential regular sense extension.

Table 1: Extracted relations for patterns

Hyper1	**Hyper1 gloss**	**Hyper2**	**Hyper2 gloss**	**Relation**	**Total Number of Participating Words**
fruit tree	tree bearing edible fruit	edible fruit	edible reproductive body of a seed plant especially one having sweet flesh	bear	88 e.g. *breadfruit, hog plum, grapefruit*
fruit tree	tree bearing edible fruit	edible fruit	edible reproductive body of a seed plant especially one having sweet flesh	cultivated for	88
fruit tree	tree bearing edible fruit	edible fruit	edible reproductive body of a seed plant especially one having sweet flesh	produce	88
covering	a natural object that covers or envelops; "the fox was flushed from its cover"	fabric	something made by weaving or felting or knitting or crocheting natural or synthetic fibers	made from	5 e.g. *acrylic, canopy, lining*
equipment	an artifact needed for an undertaking or to perform a service	game	a contest with rules to determine a winner; "you need four people to play this game"	used in	6 e.g. *football, handball, baseball*
person	a human being; "there was too much for one person to do"	job	the occupation for which you are paid; "he is looking for a job"; "a lot of people are out of work"	hold	27 e.g. *cabinet minister, PM, Treasury Secretary*
person	a human being; "there was too much for one person to do"	language	a systematic means of communicating by the use of sounds or conventional symbols; "he taught foreign languages"; "the language introduced is standard throughout the text"; "the speed with which a program can be executed depends on the language in which it	speak	257 e.g. *Tatar, Assyrian, Hopi, Punjabi*

Table 2: Topical relations of *music* (an artistic form of auditory communication incorporating instrumental or vocal tones in a structured and continuous manner)

Hypernym	**Gloss**	**Subsumed words**
social relation	a relation between living organisms; esp between people.	bass,canto,fanfare,homophony,line,obbligato, obligato,overture,resolution,rondeau,signature, Star-Spangled banner,statement,theme,voice
psychological feature	a feature of the mental life of a living organism	idea,line,macumba,melody,motif,motive,resolution,theme, variation
person	a human being.	Bach,bass,Beethoven,Brahms,Chopin,Handel, Hare Krishna,Haydn,Mozart,Stravinsky,voice, Wagner
performer	an entertainer who performs a dramatic or musical work for an audience	Bach,bass,Chopin,voice
instrumentality	an artifact (or system of artifacts) that is instrumental in accomplishing some end	bass,hornpipe,line

The various extraction processes described in this paper show the first steps towards an incremental process of adding and integrating implicit lexical knowledge into an existing explicit knowledge structure. These additional knowledge frames have not been built on the basis of an assumption of semantic completeness of WordNet, but on the assumption that WordNet based regular polysemic patterns are indicative of many instances of RP.

The incrementally growing knowledge base will provide an increasing amount of background knowledge for the text comprehension process, against which the textual data will be interpreted. This additional interpretation will be incor-

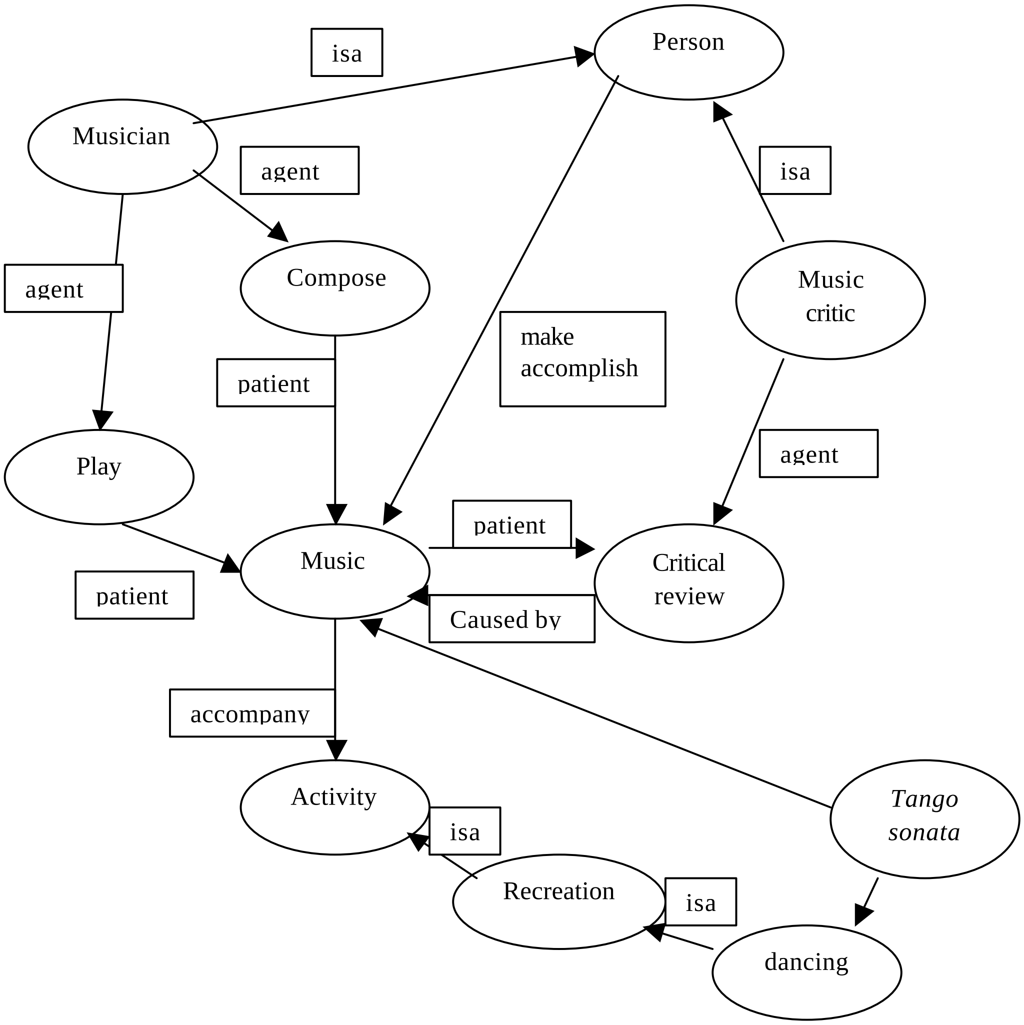

Figure 3: Extended Frame for ‘music’

porated into the lexicon, and become background for new texts. This will lead to an incremental reduction of the lexicographic burden necessary to create detailed, task-specific lexicons for applications such as information extraction. In future, semantic information from other, heterogeneous resources can be amalgamated into this single knowledge base, once correspondences between the resources have been established.

6 References

Apresjan, J. (1973), *Regular Polysemy*

In: Linguistics 142, pp. 5-32

Buitelaar, P (1998)

Corelex: Systematic Polysemy and Underspecification,

Ph.D., Department of Computer Science, Brandeis University, Boston, U.S.A.

Fellbaum, Christiane (ed.) (1998), *WordNet: An Electronic Lexical Databas*e.

Cambridge, Mass.: MIT Press.

Fillmore, C (1977), *Scenes and frames semantics.*

In: Zampolli, A (ed.) Linguistic structures processing.

Benjamins, Amsterdam, The Netherlands, pp. 55-81.

Fillmore, C. and Atkins, S. (1992), *Towards a Frame-based Organization of the Lexicon: the Semantics of RISK and its Neighbors.*

In: Lehrer, A. and Kittay, E. (Eds.) Frames, Fields and Contrasts: New Essays in Semantics and Lexical Organization, Hillsdale: Lawrence Erlbaum

Gonzalo, J., Verdejo, F, Chugur, I. and Cigarrán, J. (1998), *Indexing with WordNet synsets can improve text retrieval*

In *ACL/COLING Workshop on Usage of WordNet for Natural Language Processing*.

Lakoff, G. (1987), *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: The University of Chicago Press.

Minsky, M. (1975), *A Framework for Representing Knowledge*.

In: Winston, P.H. (Ed.), The Psychology pof Computer Vision, New York: McGraw-Hill, pp. 211-277.

Peters, W., *Self-enriching Properties of Wordnet: Relationships between Word Senses,* Proceedings of LREC 2002, Gran Canaria, 2002

Peters, W. and Wilks, Y., *Data-driven Detection of Figurative Language Use in Electronic Language Resources,*

Metaphor and Symbol Vol. 18, no. 3, pp. 161-175, 2003

Volk, M., Ripplinger, B., Vintar, S., Buitelaar, P., Raileanu, D. and Sacaleanu, B. (2002), *Semantic Annotation for Concept-Based Cross-Language Medical Information Retrieval*.

In: International Journal of Medical Informatics, Volume 67:1-3

Vossen, P.(1998), Introduction to EuroWordNet.

In: Nancy Ide, N., Greenstein, D. and Vossen, P. (eds), Special Issue on EuroWordNet. Computers and the Humanities, Volume 32, Nos. 2-3 1998. 73-89.