Lexicography in an Interlingual Ontology: An Introduction to EuroWordNet

Peter Jansen
University of Waterloo
pajansen@uwaterloo.ca

EuroWordNet is a multilingual lexical database constructed in the wake of WordNet. The ontological structure of the language-dependent layers, analogous to individual WordNets, through the semantic space of the interlingual index and abstract framework of the top level ontologies are examined. The semantic nature of the interlingual lexicon is examined as it applies to Gruber's principles for the design of ontologies. Benefits of EuroWordNet's design are highlighted.

WordNet was originally proposed by Miller (1990) in an experiment to test an implementation of a model of lexical organization. Up to this point, ontological databases had been particularly small. WordNet was intended as a test of ontology design on a scale far larger than any existing lexical database, progressively incorporating the English lexicon into a large semantic database.

The lexicon is the name given to a linguistic resource that contains our knowledge of words, including semantic data for each word or concept expressed. Concepts can be more then a single word − they can include compounds, such as 'high school', collocations, such as 'best friend', idiomatic phrases, as in 'keep in touch' or 'being under the weather', and, finally, phrasal verbs, as 'taking back' a book to the library, or 'putting on' your shoes before you go to school. Compounds, collocations, idiomatic phrases, and phrasal verbs extend the idea of storing words in the lexicon to storing conceptual information that may not have a lexical representation using a single word. WordNet currently represents a large portion of the English lexicon, consisting of over 115,000 concepts.

WordNet's organizational units are these concepts. WordNet does not contain units smaller than a word, such as phonemic or morphemic information, or larger units such as frames (Minsky, 1975), scripts (Schank and Abelson, 1977), schemas (Rumelhart, 1980), etc., consisting of multiple concepts. WordNet's structure resembles that of both a dictionary and a thesaurus, having qualities of both. A dictionary contains semantic and syntactic information about single words, and is organized by word spelling. A thesaurus contains semantically related words, and is organized by the general concept that a collection of words represents. WordNet contains *synsets*, which consist of a set of words or short word constructs (as discussed previously) which represent a specific concept. These synsets form the underlying structure of WordNet. In essence,

synsets are the reason WordNet is a semantically organized dictionary (Fellbaum,1998).

Initially, WordNet was designed to contain only synsets and pointers between the synsets called *relations*. As development progressed, definitions and example sentences were included with the concepts to help contrast related synsets. While WordNet is a lexical database, its definitions sometimes include encyclopedic knowledge to help define the concepts it represents (Fellbaum,1998).

Synsets describe a collection of lexical concepts that are semantically 'identical'. A synset may consist of only a single element, or it may have many elements all describing the same concept. Each element in a particular synset's list is synonymous with all other elements in that synset. For example, the synset {search, lookup} represents the concept of checking to see if something has a specific property. In this context, 'search' and 'lookup' are both semantically equivalent. For cases where a single word has multiple meanings (a *polysemous* word), multiple separate and potentially unrelated synsets will contain the same word.

Synsets are interconnected by relations. Relations in WordNet express simple relationships between synsets. These relations include subclass and superclass relationships (hyponymy and hypernymy), part-of / has-a relationships (meronymy and holonymy), and the antonymy, or opposite, relationship. The concept network can be traversed using these relations, and from one synset a set of relations open a meaningful path to be explored, allowing simple inferencing to take place.

WordNet consists of four distinct semantic networks, one each for nouns, verbs, adjectives, and adverbs (Fellbaum,1998). This design simplifies the network design, as each word class has different semantic relations. For instance, verbs have a relation called troponymy (Fellbaum,1998), which expresses a particular manner of doing something. Both nouns and verbs can be organized hierarchally. *Unique beginners* are noun synsets at the top of a lexicon's hierarchy (Miller, 1998). These include abstract concepts such as abstraction, possession, processes, and states. These unique beginners can serve as a conceptual base for building the semantic network from the most abstract concept towards less abstract, specific concepts and instantiations.

EuroWordNet

WordNet was designed to be used to represent English words and lexical concepts. The EuroWordNet project (Vossen, Díez-Orzas, Peters, 1997; EuroWordNet, 2001), completed in 1999, set out to create a multilingual lexical database relating conceptual information amoung a number of European languages, and to establish a common framework that would allow new languages to be incorporated. At its completion, EuroWordNet combined the Czech, Dutch, Estonian, Italian, French, German, and Spanish languages, and, since the project's end, a number of additional languages have been developed to its specification, including Swedish and Russian (EuroWordNet, 2001).

The EuroWordNet team examined a number of designs for their multilingual system (Vossen et al., 1997). One of the more expansive approaches considered was to map concepts in one language to concepts in each of the other languages. In this way, if the multilingual database consisted of three languages, six different interlingual conceptual mappings would need to exist (one from each language to each other language). For instance, a potential set of mappings might be English to French, French to English, English to German, German to English, French to German, and German to French. The effort required to add new languages in this system becomes extremely large as the number of languages increases. The potential advantage of this method however would be the tailored translations between languages, which may make interlingual mappings more precise.

The actual design used by the EuroWordNet team requires less computational resources, but with some added advantages and disadvantages. The design is as follows. The database is organized into three main layers: the language-dependent layer, the language-independent layer, and the top-layer and domain ontologies. The language-dependent layer consists of a WordNet structured similarly to the English WordNet, containing the concepts for one specific language. Each language-dependent layer is in essence a WordNet of its own for a specific language. These multiple WordNets are then connected to a language-independent lexical database. This database, called the *interlingual index* (ILI), is a WordNet of its own, but unlike a language-dependent WordNet, its synsets link to the synsets of other language-dependent WordNets. The synsets contained in the ILI represent language-independent concepts, free of the lexical constraints of any one language. In this way, the concepts represented in different languages are cross-lingually linked together, and a concept specified in any one language can be translated into any other language connected to the ILI.

The synsets were developed hierarchically between languages by first identifying common 'base concepts', or concepts that were common to all languages, and beginning the database development from these base concepts. Thirty representative synsets were selected by all language-specific developers, of which 24 are noun synsets, and six are verb synsets. In situations where the language-specific developers identified more base concepts, the concepts were further abstracted to the common set of base concepts. In instances where a base concept isn't lexically represented in a language, a close representation is used.

The base concepts are organized into a top-level ontology where the base concepts are hierarchically extended to include closely related hyponyms. The base concepts are divided into two categories in the top-level ontology: high order entities, and first order entities. High order entities are abstract concepts and include events, processes, relations, properties, and states. First order entities are material objects and perceivable quantities. The top layer of EuroWordNet also contains the domain hierarchy ontology, which allows synsets in the interlingual index to be mapped directly to categorical descriptions, for instance, animal, vertebrate, invertebrate, plant, or clothing. The top-level ontology labels and the domain labels have equivalence relations to synsets in the ILI. This design feature is useful in instances where language-independent but domain-specific ontologies designed for a specific task may be required. Linking to a domain ontology may also help select more generic (further away) or more specific (closer) concepts in interlingual translation (Vossen et al., 1997, p. 2).

The ILI contains six different relations specific to the layer's development (Vossen et al., 1997, p. 3-7). These relations are useful in situations where languages don't map well to each other. Some languages have concepts which are not lexicalized in others. For instance, the English word 'head' can refer to any head, but in Dutch there are different words to express either 'human head' or 'animal head' (Vossen et al., 1997, p. 4). This situation represents one of these ILI relationships, HAS_EQ_HYPERONYM, when a concept exists in one language which is more specific than an existing synset in the ILI. Other relations include HAS_EQ_HYPONYM, where a concept is too general for an existing synset and is mapped to a more specific synset, and HAS_EQ_SYNONYM, where concepts in the ILI are synonymous or identical to each other.

A number of desiderata were introduced by Gruber (1993) to help guide the development of and serve in evaluating ontologies (Gomez-Perez, 2003). These guiding principles, which we shall examine as they apply to EuroWordNet, include coherence, clarity, extendibility, minimal encoding bias, and minimal ontological commitments.

The principal of coherence states that

inferences created through the use of the ontology should not lead to contradictions. A contradiction means that the ontology contains incoherent information. The possible sources of contradiction in EuroWordNet could include situations where closely related concepts are independently categorized, or categorized by different developers and, as a result, synsets in the ILI may actually have both hypernym and hyponym relations to another specific synset. This type of error has been minimized at the higher levels of the ontology by using a common set of base concepts to develop each language-dependent WordNet. Automated searches for synsets that contain subclass-of and superclass-of relations to another synset could be used to find such issues, then either the user or automated inferencing (perhaps selecting the most common hierarchical derivation found between the languages) could correct the incoherence.

Clarity is the principle that terms should be effectively communicated. In terms of the structure of individual WordNets, definitions to help differentiate semantically similar synsets should be clear. The top-level ontology should also clearly express each base concept. Due to the highly abstract nature of these concepts, the base concepts may be best illustrated through elaborating subordinate nodes, perhaps through multiple levels. Clarity would not seem to apply to the interlingual index, as the concepts it contains are purely conceptual and must be interpreted into a language in order to be linguistically perceived.

Extendibility is the guiding principle behind the design of EuroWordNet. The specifications allow additional languages to be mapped into EuroWordNet's structure with a minimum of effort. The principal of extendibility states that an ontology should be able to support the addition of hyponyms to existing concepts without modifying pre-existing concepts. The use of a common base-concept ontology developed through examining commonalities between multiple languages suggests that violations of the extendibility principle should be kept to a minimum, and would likely occur in an active revision of the top-level ontologies while adding additional languages. The dynamic, network-like nature of synsets should allow complete extendibility beyond the top-level ontology.

The minimal encoding bias states that concepts should be defined at a 'knowledge level' and should not be dependent on a symbolic level of encoding. This principle alludes to the use of the common top-level ontology using common base concepts in the development of the language-dependent WordNets. In this way, concepts in all languages are build upon this highly abstract layer, which should minimize the bias that could exist if the language-dependent ontologies were built upon unique top-level ontologies.

Finally, the notion of minimal ontological commitment signifies minimizing specificity of information that could exist in different formats. This is an especially important consideration in a cross-cultural, interlingual database. Examples of this bias could include measurements such as dates, spans of time, distances, and intensities. The synset nature of EuroWordNet elegantly expresses the spirit behind this principle by expressing information semantically. Problems where encoding biases may occur could include the synset definitions in each language, which may state each language's specific method of interpreting some concept such as measure or quantity.

EuroWordNet attempts to incorporate a large portion of the semantic lexicons of multiple European languages in a common framework. The design of this framework is flexible enough to allow the relatively easy addition of new languages, and scales tractably both in terms of computational resources required to process the lexical database, and the work required to create new linguistic databases and connect them with EuroWordNet. The semantic nature of synsets embodies many of Gruber's (1993) principles of ontological development, and combined with systems for semantic disambiguation, could form

an impressive interlingual translation system. While the project was officially completed in 1999, the specification continues to be used and nearly three times the number of languages originally supported have individual WordNets developed and can be linked to EuroWordNet's interlingual index. ■

References

EuroWordNet. (2001). http://www.illc.uva.nl/EuroWordNet

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press. 1-12.

Gomez-Perez, A., Corcho, O., Fernandez-Lopez, M. (2003). Ontological Engineering: With Examples from the Areas of Knowledge Management, E-Commerce, and the Semantic Web. Springer Verlag.

Gruber, T. (1993). Towards Principles for the Design of Ontologies Used for Knowledge Sharing. Technical Report KSL93-04, Stanford University, Knowledge Systems Laboratory.

Miller, G. A. (1990). WordNet: An On-line Lexical Database. International Journal of Lexicography, 3, 235-312.

Miller, G. A. (1998). Nouns in WordNet. In: WordNet: An Electronic Lexical Database. Fellbaum, C. (Ed.). Cambridge, MA: MIT Press. 23-46.

Minsky, M. (1975). A Framework for Representing Knowledge. In P. H. Winston (Ed.), The Psychology of Computer Vision 211-277. New York: McGraw-Hill.

Rumelhart, D.E. (1980). Schemata: The Building Blocks of Cognition. In R.J. Spiro, B.Bruce, & W.F. Brewer (eds.), Theoretical Issues in Reading and Comprehension. Hillsdale, NJ: Erlbaum.

Schank, R. C., & Abelson, R. P. (1977). Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures. Hillsdale, NJ: Lawrence Erlbaum.

Vossen, P., P. Díez-Orzas, W. Peters. (1997). The Multilingual Design of EuroWordNet. In: P. Vossen, N. Calzolari, G. Adriaens, A. Sanfilippo, Y. Wilks (eds.) Proceedings of the ACL/EACL-97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for Natural Language Processing Applications, Madrid, July 1997.