Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System

Anita BURGUN
National Library of Medicine
8600 Rockville Pike
Bethesda, MD, 20894
burgun@nlm.nih.gov

Olivier BODENREIDER
National Library of Medicine
8600 Rockville Pike
Bethesda, MD, 20894
olivier@nlm.nih.gov

Abstract

Objectives: The objective of this study is to compare how a general terminological system (WordNet) and a domain-specific one (UMLS) represent linguistic and knowledge phenomena at three different levels: terms, concepts, and semantic classes. **Methods**: For one general class (ANIMAL) and one domain-specific class (HEALTH DISORDER), the set of concepts corresponding to the class was established. Then, for each semantic class, the corresponding terms were mapped from one system to the other, both ways. **Results**: Only 2% of the domain-specific concepts from UMLS were found in WordNet, but 83% of the domain-specific concepts from WordNet were found in the UMLS. Concept overlap between the two systems varies from 48% to 97%. **Discussion**: Missing terms in both systems are discussed, as well as granularity and knowledge organization issues.

Introduction

The Unified Medical Language System® (UMLS®) has been developed and maintained by the National Library of Medicine since 1990. It is intended to help health professionals and researchers use biomedical information from different sources (Lindberg, Humphreys, & McCray, 1993). The current version (2001) integrates about 800,000 concepts from more than fifty families of vocabularies such as the International Classification of Diseases or Medical Subject Headings (UMLS, 2001). While the structure of each source vocabulary is preserved, terms that are equivalent in meaning are clustered into a unique concept. Furthermore, interconcept relationships, either inherited from the source vocabularies or specifically generated, give the UMLS Metathesaurus additional semantic structure. Each Metathesaurus concept is assigned to at least one of the 134 semantic types from the Semantic Network, providing each concept a categorization that is independent from its relationships to other concepts, as detailed below.

WordNet®, an electronic lexical database, has been developed and maintained at Princeton University since 1985 (Fellbaum, 1999). Sets of synonymous terms, or synsets, constitute its basic organization. The current version (1.6) integrates about 100,000 synsets. Several types of relations between synsets are recorded in WordNet, including hyponymy and meronymy.

The following differences can be pointed out between WordNet and the UMLS.

1) **Terms**: Although both systems have terms, WordNet only records their canonical form, while the UMLS records all strings provided by medical vocabularies for a given term, including inflectional variants, case and hyphen variants, and variants related to the presence of terminological modifiers that do not affect the meaning (e.g., “not otherwise specified”). In addition, the UMLS integrates the translation of some vocabularies in many languages. Only English UMLS terms are used in this study.

2) **Concepts**: A cluster of synonymous terms is called a *synset* in WordNet and a *concept* in the UMLS. Beside the difference in their names, WordNet synsets and UMLS concepts are

Proceedings of NAACL'2001 Workshop, "WordNet and Other Lexical Resources:
Applications, Extensions and Customizations" 2001:77-82.

functionally equivalent and have a unique identifier in each system. For example, the meaning "prostate" is represented by the WordNet synset *04187344*, cluster of the synonymous terms "prostate" and "prostatic gland", and by the UMLS concept *C0033572*, cluster of the synonymous terms "Prostate", "Prostatic gland", "Prostate gland" and "Glandula prostatica".

3) **Semantic classes**: In the UMLS, interconcept relationships are not always defined with precision. Hierarchical relationships, though generally hyponymic, may also be meronymic or reflect whatever principle a given vocabulary uses to define hierarchies. For this reason, it is difficult to rely on hierarchical relationships in the UMLS for establishing semantic classes (i.e. the sets of concepts corresponding to a given semantic category), because the children of a concept are not necessarily all its hyponyms. Although WordNet and the UMLS have different structures, semantic classes can be compared in the two environments. In WordNet, we define a class as the set of all hyponyms of a given synset. In the UMLS, a class is defined as the set of concepts that are assigned to a given semantic type. It is expected that, for a general category (e.g., ANIMAL), there will be a large overlap between semantic classes in WordNet, representing a general terminological system, and in the UMLS, representing a domain-specific one. Conversely, for a semantic category central to the medical domain (e.g., HEALTH DISORDER), the class in WordNet is expected to be essentially included in the equivalent class in the UMLS, since the degree of specialization (or granularity) of a domain-specific terminology is higher.

Several other studies attempted to merge linguistic or knowledge resources with WordNet, but in a domain and with a perspective different from ours. (Kwong, 1998) aligned WordNet with other general lexical resources such as Roget's Thesaurus, and showed how a combination of resources may be helpful for natural language processing applications. (O'Sullivan, McElligott, & Sutcliffe, 1995) mapped WordNet with an ontology specific to the domain of computer science, and (Kiryakov & Simov, 2000) compared the upper-level ontologies in EuroWordNet and Cyc. The objective of our study is to compare how a general terminological system (WordNet) and a domain-specific one (UMLS) represent linguistic and knowledge phenomena at three different levels: **terms** (the symbols), **concepts** (the clusters of terms corresponding to a given meaning), and **semantic classes** (the sets of concepts corresponding to a given semantic category).

1 Methods

We focused on two semantic classes for comparing WordNet and the UMLS: ANIMAL, a general class, and HEALTH DISORDER, typical of the medical domain. We first established the classes by selecting the corresponding WordNet synsets and UMLS concepts. We then mapped WordNet terms to the UMLS and UMLS terms to WordNet. Finally, we compared the terms, concepts and classes in both systems.

1.1 Establishing the semantic classes

In WordNet, the semantic classes were established by using the hyponymic relation, starting with a given high-level synset. A class consists of this high-level synset and all its hyponyms.

In the UMLS, it was possible to take advantage of the semantic categorization provided for each Metathesaurus concept for establishing the semantic classes. A class consists of all the concepts that are assigned to a given semantic type.

In the case of the HEALTH DISORDER class, several high-level synsets in WordNet and several semantic types in the UMLS were needed for seeding such a broad class. Details about the constitution of the classes are provided in Table 1.

1.2 Mapping WordNet terms to the UMLS

Each term from a WordNet semantic class was mapped to the UMLS through the Knowledge Source Server[1] (UMLS, 2001). If a WordNet term did not exactly match to the UMLS, a normalized match was attempted. Normalization addresses mostly inflection, case and hyphen variation, and word order variation.

[1] umlsks.nlm.nih.gov

	WordNet	UMLS
ANIMAL	The synset *Animal* and all its hyponyms	The UMLS concepts assigned to the semantic type *Animal*, or any of its subtypes *(Invertebrate, Vertebrate, Amphibian, Bird, Fish, Reptile, Mammal, Human)*
HEALTH DISORDER	The union of the following synsets and all their hyponyms: • *Symptom* • *Ill Health* • *Disorder (sense 1)* • *Mental retardation* • *Mental Illness* • *Defect (sense 1)* • *Abnormalcy*	The UMLS concepts assigned to any of the following semantic types: • *Anatomical Abnormality* • *Congenital Abnormality* • *Acquired Abnormality* • *Finding* • *Sign or Symptom* • *Pathologic Function* • *Disease or Syndrome* • *Mental or Behavioral Dysfunction* • *Neoplastic Process* • *Cell or Molecular Dysfunction* • *Experimental Model of Disease* • *Injury or Poisoning*

Table 1 – Definition of the semantic classes in WordNet and the UMLS.

For each term of each WordNet synset, the following elements were recorded: the presence of an equivalent lexical item in the UMLS, the matching method (exact match/ normalization), and information about the UMLS concept mapped to, especially whether this concept belongs to one of the UMLS semantic classes of interest. One WordNet term may map to several UMLS concepts. For example, "allograft" maps to *C0522536* (the procedure of transplanting) as well as to *C0085769* (the transplanted organ or tissue).

1.3 Mapping UMLS terms to WordNet

Each term from a UMLS semantic class was mapped to WordNet, using the standard function *wn*. Since *wn* ignores parenthetical expressions, parentheses were removed from UMLS terms prior to mapping them to WordNet. The vast majority of biomedical terms are noun phrases in which the head noun is modified either by an adjective, another noun or a prepositional phrase. Therefore, we restricted the mapping of UMLS terms to nouns in WordNet.

For each string of each UMLS concept, the following elements were recorded: the presence of an equivalent lexical item in WordNet, and information about the WordNet synset mapped to, especially whether this synset belongs to one of the WordNet semantic classes of interest. One UMLS term may map to several WordNet synsets. For example, "arteries" maps to *04137243* (the blood vessel) and *02213687* (the street).

1.4 Comparing terms, concepts and semantic classes

Terms were considered equivalent if they mapped successfully by the method referenced above.

Concepts were determined to be equivalent if at least one term from the WordNet synset was equivalent to at least one term from the UMLS concept. We call *total mapping* the situation where all the terms of a UMLS concept or WordNet synset are found in the other system. In case of multiple mapping of one term on one side to several terms on the other side, the mapping was considered relevant if one of the terms mapped to belonged to the semantic class of interest.

The comparison of **semantic classes** was based on the overlap between sets of concepts or synsets in both systems for a given class. Let us consider an element, i.e. synset or a concept, from the class C in the system S. C' is the class that corresponds to C in the system S'. For a given class C, the overlap of concepts between the two systems is given by the ratio between the number of elements that belong to C and have at least an equivalent that belongs to C' and the number of elements that belong to C that have an equivalent in S'. For example, this ratio represents the conditional probability that a WordNet concept from the ANIMAL class belongs to the UMLS ANIMAL class, given that this concept belongs to the UMLS.

		ANIMAL		HEALTH DISORDER	
		Nb. of synsets	%	Nb.of synsets	%
Original set of synsets		3,984		1,379	
Coverage	Synsets found in the UMLS	2,046	51%	1,144	83%
	Synsets not found in the UMLS	1,938	49%	235	17%
When found	Concept overlap between WordNet and the UMLS (synsets found in the equivalent UMLS class)	1,919	94%	1,112	97%
	Total mapping of concepts (the synset is found and all its terms are found)	1,012	49%	979	86%

Table 2 – Mapping of WordNet terms to UMLS.

		ANIMAL		HEALTH DISORDER	
		Nb. of concepts	%	Nb.of concepts	%
Original set of concepts		11,634		143,991	
Coverage	Concepts found in WordNet	2,154	19%	2,639	2%
	Concepts not found in WordNet	9,480	81%	141,352	98%
When found	Concept overlap between the UMLS and WordNet (concepts found in the equivalent WordNet class)	1,582	73%	1,257	48%
	Total mapping of concepts (the concept is found and all its terms are found)	973	45%	413	16%

Table 3 – Mapping of UMLS terms to WordNet.

2 Results

2.1 Mapping WordNet terms to the UMLS

ANIMAL. 3,984 WordNet synsets from the ANIMAL class and 7,961 WordNet terms were mapped to the UMLS. Among them, 2,046 synsets (51%) and 2,895 terms (36%) were mapped successfully (Table 2). Exact match was involved in 84% of successful mapping. Among WordNet synsets found in the UMLS, the percentage of synsets found in the equivalent class was 94%.

HEALTH DISORDER. 1,379 WordNet synsets representing the HEALTH DISORDER class and 2,194 WordNet terms were mapped to the UMLS. Among them, 1,144 synsets (83%) and 1,699 terms (77%) were mapped successfully (Table 3). Exact match was involved in 95% of successful mapping. 80 WordNet synsets not found in the UMLS corresponded to plant diseases, mostly outside the biomedical domain covered by the UMLS. Among WordNet synsets found in the UMLS, the percentage of synsets found in the equivalent class was 97%.

As suggested by the differences observed in mapping between concepts and terms, the rate of total mapping confirmed that, for a given class, the overlap between WordNet and the UMLS is higher for concepts than for terms. Finally, for 165 WordNet HEALTH DISORDER synsets with an equivalent in the UMLS, some terms are not present in the UMLS.

2.2 Mapping UMLS terms to WordNet

ANIMAL. 11,634 UMLS concepts from the ANIMAL class were mapped to WordNet. 19% were found in WordNet. Among UMLS concepts found in WordNet, the percentage of concepts found in the equivalent class was 73%.

HEALTH DISORDER. More than 140,000 concepts representing the HEALTH DISORDER class in the UMLS were mapped to WordNet. 2% were found in WordNet. Among UMLS concepts found in WordNet, the percentage of concepts found in the equivalent class was 48%.

3 Discussion

3.1 Missing terms

When mapping WordNet to the UMLS, the concept coverage for the class HEALTH DISORDER is quite good (83%). However, the coverage of terms is lower (77%), and only 86% of the synsets are found in the UMLS with all their terms. Practically, it means that terms represented in WordNet are sometimes absent from the medical vocabularies integrated in the UMLS. These terms are essentially lay terms for disorders. For example, a synonym of "infectious mononucleosis" in WordNet is "kissing disease", which does not exist in the UMLS. This phenomenon is of potential interest for augmenting lay terminology in the UMLS, with applications in consumer health projects, for example.

When mapping the UMLS to WordNet, the proportion of total mapping of concepts is very low, meaning that the UMLS contains many terms that are not represented in WordNet, even when concepts exist in both systems. Part of those terms belong to the vocabulary of a specialized domain. For example, "coal pneumoconiosis" is a specialized medical synonym for "anthracosis/ black lung disease". Other terms are missing because of terminology-specific modification in a particular vocabulary. Examples of such features include markers for underspecification (e.g., "Generalized epilepsy, *not otherwise specified*", used when the etiology or the clinical form of the disease is not precisely known) and classification-specific terms (e.g., "Generalized epilepsy, *without mention of intractable epilepsy*", as opposed to "Generalized epilepsy, *with intractable epilepsy*"). Inverted terms, created for helping humans search terms in a list, are also typically absent from WordNet (e.g., "Epilepsy, generalized").

3.2 Granularity

Domain terminologies require precise distinction among concepts. For example, the two terms "grand mal epilepsy" and "generalized epilepsy" are synonymous in WordNet, while there are two distinct concepts in the UMLS. Medically, "grand mal epilepsy", also called "tonico-clonic epilepsy", is a kind of "generalized epilepsy", along with "tonic epilepsy", among others. Therefore, technically, "generalized epilepsy" and "grand mal epilepsy" are better represented in hierarchical relationship as in the UMLS, than as synonyms as in WordNet. Practically, however, the semantic distance between these two concepts is probably small enough for them to be used interchangeably in lay usage.

Even within a specialized terminology, the choice between clustering two terms into a unique concept or creating a distinct concept for each of them is sometimes arbitrary or driven by pragmatic considerations. From a theoretical perspective, some pairs of terms should be considered plesionyms rather than synonyms (Cruse, 1986).

3.3 Knowledge Organization

Although found in both WordNet and the UMLS, a concept may be categorized differently in the two systems. The following two situations may occur.

1) In the target system, the concept belongs to a semantic class that is different from that in the source system. For example, the UMLS does not isolate biological taxons from animals. Therefore, "Cetacea" is categorized as an ANIMAL in the UMLS while it is a hyponym of "Taxonomic Group" in WordNet. This phenomenon reflects differences in knowledge organization. Similarly, it is not surprising that the WordNet synsets "Carnivora" (hyponym of "Taxonomic Group") and "carnivore" (hyponym of "Animal") are clustered into the same UMLS concept.

2) Within the same broad semantic class HEALTH DISORDER, formed from the union of several categories, concepts may be categorized differently by WordNet and the UMLS. Even if the definition of the categories looks similar in the two systems, the list of concepts found under these categories may be different. For example, "Symptom" has equivalent definitions in WordNet, where its is 'any sensation or change in bodily function that is experienced by a patient and is associated with a particular disease', and in the UMLS, where "Sign or Symptom" is 'an

observable manifestation of a disease or condition based on clinical judgment, or a manifestation of a disease or condition which is experienced by the patient and reported as a subjective observation'. This semantic similarity leads to a certain degree of overlap. For example "cyanosis" correctly belongs to both the hyponyms of WordNet "Symptom" and the set of UMLS concepts that are assigned the semantic type "Sign or Symptom". However, "Symptom" in WordNet is also a hypernym of "encephalitis", "sinusitis", "tennis elbow", and numerous other conditions that are assigned the semantic type "Disease or Syndrome" in the UMLS.

3.4 Hyponymy vs. Categorization

Some 1,382 (52%) HEALTH DISORDER concepts in the UMLS were found in WordNet, but outside the HEALTH DISORDER class as defined in Table 1. For example, "bronchospasm" is the spasmodic contraction of the smooth muscle of the bronchi, that occurs in asthma, among other diseases. The UMLS categorizes it both as a "Disease or Syndrome" and a "Sign or Symptom". In WordNet, however, none of the hypernyms of "bronchospasm" belongs to the class HEALTH DISORDER. The emphasis is put on the general physical mechanism involved in the spasm (spasm, constriction, squeeze, […], action, act), rather than on its pathological aspects. As a consequence, WordNet allows for "bronchospasm" to inherit features of its hypernyms (e.g., "constriction"), which is not possible in this case in the UMLS. However, whatever the principles behind the organization of concepts in the UMLS, the categorization of the concepts according to semantic types from the Semantic Network enables users to consistently select sets of concepts that belong to a given category.

Conclusion

This study compares WordNet and the UMLS in their representation of linguistic and knowledge phenomena at the levels of terms, concepts and semantic classes. While there is no major difference in the representation of concepts (clusters of terms in both systems), the representation of terms and classes shows more differences. For terms, the UMLS records the variability of the lexical forms encountered in the source vocabularies, while WordNet only records the canonical form. For classes, in the UMLS, 134 high-level categories provide an additional semantic structure, offering a simple way to categorize the concepts.

Our methodology provides a way to integrate lay vocabulary from WordNet into a medical thesaurus. Conversely, this method can be used for extending WordNet with vocabulary from a specialized thesaurus.

Acknowledgements

This research was supported in part by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine.

References

Cruse, D. A. (1986). *Lexical semantics*. Cambridge; New York: Cambridge University Press.

Fellbaum, C. (Ed.). (1999). *WordNet: An electronic lexical database*. Cambridge, Massachusets: MIT Press.

Kiryakov, A., & Simov, K. I. (2000). *Mapping of EuroWordNet Top Ontology to Upper Cyc Ontology*. Paper presented at the EKAW 2000 Workshop on "Ontologies and Text", Juan-les-Pins, French Riviera, Oct. 2-6, 2000.

Kwong, O. Y. (1998). *Aligning WordNet with additional lexical resources.* Paper presented at the COLING-ACL98 Workshop on "Usage of WordNet in Natural Language Processing Systems", Montreal, Canada.

Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The Unified Medical Language System. *Methods Inf Med, 32*(4), 281-291.

O'Sullivan, D., McElligott, A., & Sutcliffe, R. F. E. (1995). *Augmenting the Princeton WordNet with a Domain Specific Ontology.* Paper presented at the IJCAI'95 Workshop on "Basic Ontological Issues in Knowledge Sharing", Montreal, Canada, August 19-21, 1995.

UMLS. (2001). *UMLS Knowledge Sources* (12th ed.). Bethesda (MD): National Library of Medicine.