Adapting WordNet to the Medical Domain using Lexicosyntactic Patterns in the Ohsumed Corpus.

Toumouh A.[1], Lehireche A.[1], Widdows D.[2], Malki M.[1]

[1]Computer Science Departement,
University Djillali Liabès.
Sidi Bel Abbes, Algeria
email toumouh@gmail.com

[2]MAYA Design INC
Pittsburgh, Pennsylvania, USA
email widdows@maya.com

Abstract

Ontologies are widely used in several areas with applications including knowledge Management, Web commerce and electronic business. An ontology provides a consensus of concept specifications for a specific domain shared by a group of people. In this paper we deal with Ontology Learning, specifically we aim to adapt the WordNet ontology, a general source of lexical knowledge, to the medical domain. We use for this task a combination of lexico–syntactic pattern, mainly conjunctions of the form "Noun_CJC_Noun", where CJC can be {and, or, but}. Pairs of words extracted in this fashion are compared to find their similarity in the WordNet noun hierarchy, using a form of the Resnik similarity method. Large scale experiments were conducted by extracting many such pairs of nouns from the Ohsumed corpus and mapping them into WordNet. For a noun pattern like "A or B" we find the lowest common ancestor of A and B by using the hypernym and hyponym links. This enables us to keep the appropriate medical sense of the two words A and B.

Key words

Ontology Learning, WordNet, Pattern, Hyponymy, Hypernymy, Synonymy, Similarity, Lowest Common Ancestor.

1. INTRODUCTION

Ontologies are widely used in several areas with applications like knowledge management, Web services and electronic business. Ontologies provide a consensus of concept specifications for a specific domain, shared by a group of people or systems. Due to the increasing abundance of specialist terminology in almost every field of human knowledge, each domain needs ontological resources tailored to its particular needs.

The challenge to equip each domain with a suitable ontology has given rise to a new research area: "Ontology Learning". It is an emerging field aimed at reducing the manual effort for engineering and managing domain ontologies. The reuse of existing ontologies that have been built manually will make the process of constructing ontologies less consuming in cost and time. For this reason we propose that one kind of Ontology Learning is to adapt general ontologies to specific domains. This requires that at least two problems be addressed:

i. **Enrichment.** New domain-specific terms need to be extracted and added to the ontology.
ii. **Pruning.** Terms from the general ontology have to be checked to see if their meaning is appropriate to the specific domain.

The ontology enrichment problem has been addressed using a variety of approaches (see for example [6], [15], [4, Ch 5]). At the time of writing, a combination of techniques has been developed, including elements of distributional analysis and pattern-based extraction.

This paper is therefore devoted to the second problem, that of "sense selection" or "pruning". General purpose ontologies tend to include many polysemous terms, i.e. ambiguous terms that have many senses, few of which are relevant to the specific domain. A generally accepted fact is that the more one restricts the domain of discourse, the more one reduces the average ambiguity of terms. For example, WordNet 2.1 gives six senses for the word *joint*, but only one of these senses ("the point of connection between two bones or elements of a skeleton (especially if it allows motion)") is usually relevant to the medical domain. To adapt the WordNet ontology to the medical domain, it would be necessary to select this sense as the "most relevant sense" from a medical point of view. This paper uses a similar idea for the purpose of ontology adaptation. In related work, Buitelaar and Scaleanu [2] tried to find medical senses of German terms, using a combination of morphological decomposition, pattern analysis and instance based learning.

The purpose of this paper is to demonstrate that it is possible to adapt a general ontology (WordNet) to a specific domain (medicine), by comparing the structure of word senses given by the general ontology with word usages in domain-specific documents. Words that are closely related to one another in documents are normally being used with meanings that are related to one another in the general ontology. For example, a document containing the phrase "doctor and nurse" is probably using *doctor* to mean *physician* rather than *learned person*. This observation was made by Resnik in [11] and used for the purpose of word-sense disambiguation.

On the basis of these ideas we aim to adapt WordNet, recognized as a general lexical ontology, to the medical domain, by comparing the senses of ambiguous words given by WordNet with the usage patterns of these words in a corpus of medical documents. We proceed by combining lexico-syntactic patterns of the form "NOUN_CJC_NOUN" (eg: A and/or B) and the Resnik similarity method [11] in order to extract from WordNet the right sense of terms belonging to the medical domain. The underlying claim is that words that are semantically similar occur with similar distributions and in similar contexts [9]. The use of the specific NOUN_CJC_NOUN pattern enables us to restrict our inferences of semantic similarity to pairs of words whose cooccurrence is especially likely to be semantically significant.

For a given word pair (whose two words are related by a CJC), we find the Lowest Common Ancestor (henceforth referred to as the LCA) by using the hypernymy relation. We then navigate back down using the hyponymy relation, in order to identify the right sense for each of the input words. This approach can be thought of as a word-sense disambiguation technique, since it allows us to choose between many senses and keep the most likely medical sense.

This paper proceeds as follows. In section 2 we give a brief outline of the WordNet ontology, and in section 3 we describe similarity, relatedness and how to use similarity for adapting WordNet. Sections 4 and 5 explain the architecture of the stages used to attempt the final results. Section 6 presents what we have obtained as results, explanations and causes of success and failures for the different cases. Section 7 gives a brief conclusion and analysis of future directions.

2. WORDNET

The Princeton WordNet [4] is a freely available, broad coverage lexical resource whose design is inspired by psycholinguistic theories of human lexical memory [9]. WordNet classifies words into four categories: nouns, verbs, adjectives and adverbs. In WordNet, each word can be associated with many senses. A word sense is identified by a set of terms called a synonym set or "synset". Each synset includes a specific concept that is defined through a gloss and eventual examples. Two kinds of relations exist in WordNet: lexical relations hold between word forms and semantic relations hold between synsets. Unlike synonymy and antonymy, which are lexical relations, hyponymy and hypernymy are semantic relations between word meanings. "Hypernym" is the generic term used to designate a whole class of specific instances. Y is a hypernym of X if X is a (kind of) Y. Hyponym is the specific term used to designate a member of a class. X is a hyponym of Y if X is a (kind of) Y. Because of this, hypernym and hyponym relations are sometimes describes as *is-a* relations. Our attention has been focused on nouns, which are organized principally by using these taxonomic relations.

The choice of WordNet for our experiments was based upon three factors:

i. WordNet is widely used and freely available, making it easy for our results to be replicated by others.
ii. There is a clear distinction between word forms and their lexical relations, and synsets and their semantic relations. This structure makes it much easier to select parts of WordNet based upon semantic considerations.
iii. Many researchers have used WordNet as a standard from which to develop other language resources, e.g. adaptations to Arabic [3] and European languages [13].

At the same time, WordNet does have its drawbacks. Being first and foremost a lexical semantic resource, WordNet is strong on relationships between words and meanings, but sometimes suboptimal for describing ontological relationships between things. For example, some world religions are classed as "psychological states", whereas others are classed as "groupings of people", and in general each religion has aspects of both these hypernyms.

3. SEMANTIC MEASURES AND WORDNET ADAPTATION

Measures of similarity quantify how much two concepts are alike, based on information contained in an *is-a* hierarchy. For example, an *automobile* might be considered more like a *boat* than a *tree*, if *automobile* and *boat* share *vehicle* as a common ancestor in an *is–a* hierarchy. It is important to note that semantic relatedness is a more general concept than *similarity*; similar entities are semantically related by virtue of belonging to a common "semantic field" (e.g. *bank–trust company*), but dissimilar entities may also be semantically related by other relationships such as meronymy (*car–wheel*) and antonymy (*hot–cold*), or just by any kind of functional relationship or frequent contextual association (*pencil–paper, penguin–Antarctica*)[1]. Resnik [11] gives the following example of relatedness and similarity: *cars* and *gasoline* would seem to be more closely related than, say, *cars* and *bicycles*, but the latter pair are certainly more similar. In this paper we focus on the *is-a* relationships that tell us when two objects are genuinely semantically similar.

Much work has been done around measuring similarity in hierarchies, for example, defining the similarity between two nodes by the length of the shortest path between them. Resnik [11] defines the similarity in taxonomy by finding the most highly specific concept that subsumes two words and gives each node its "information content". Information content, a numeric value, is a concept of information theory that defines the more informative node as being the one that occurs less frequently. In our experiments, the Resnik method of finding the LCA is used, without the notion of information content.

Given a pair of nouns extracted from a corpus (in this case, the Ohsumed medical corpus [7]), we compare their possible senses as follows:

First: We find the LCA of the two nouns in the conjunctions from the corpus then we use hyponym links in order to go down and keep the senses of the two words supposed the most appropriate to the specific domain (medical domain). This idea usually matches well because the lexico-syntactic pattern, which we have adopted, "Noun_CJC_Noun", usually signifies that the two nouns in question are similar.

Second: How we will deal with each noun having many senses in WordNet? Widdows [15] explains that the measuring of similarities becomes complicated by the appearance of ambiguity, but that this can also be used as opportunity to resolve ambiguity. For example, the word *artery* has two senses (is mapped to two synsets or concepts) in WordNet. If we encounter the phrase "artery and roadway", the sense of artery which is more semantically near the sense of roadway will be the second sense of *artery* in WordNet, and consequently the LCA will be *road* (and not *entity* which is higher in the taxonomy) reached by the unbroken path (see figure 1). On the other hand, in the example "artery and ductus arteriosus", the sense given to *artery* is the medical one (*blood vessel*). The LCA is *blood vessel* shown by the unbroken path (see figure 2).

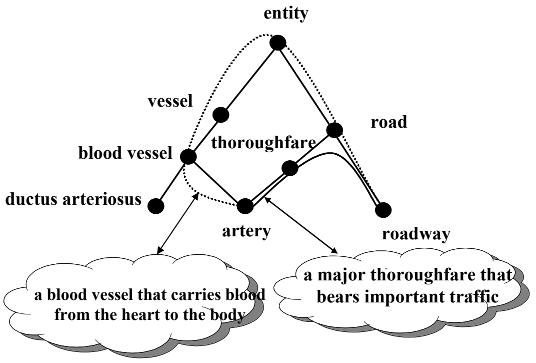
entity
vessel
road
blood vessel
thoroughfare
ductus arteriosus
artery
roadway
a blood vessel that carries blood from the heart to the body
a major thoroughfare that bears important traffic

Figure 1.The common ancestor of (artery and roadway)

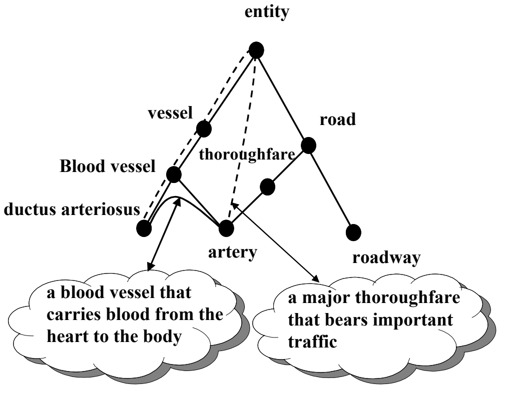
entity
vessel
road
Blood vessel
thoroughfare
ductus arteriosus
artery
roadway
a blood vessel that carries blood from the heart to the body
a major thoroughfare that bears important traffic

Figure 2. The common ancestor of (artery and ductus arteriosus)

4. THE ADOPTED APPROACH

A word may have many senses in WordNet. In practice, one often needs to measure word similarity, rather than concept similarity (see [11]). For two words w_1 and w_2 with respectively n_1 and n_2 senses in WordNet, we find for each combination, the common ancestors, and then we select the most specific (lowest) candidate. The LCA for two words w_1 and w_2 is computed as follows: a set of common ancestors is constituted by collecting the common ancestor for each one of the $n_1 * n_2$ possible combinations. The LCA of the two words w_1 and w_2 is the most specific one in this set. It is also specified as the deepest ancestor from the top node. We record A, the set of common ancestors for two senses s_1, s_2. We define S_Low_Com_Anc(s_1,s_2) as the lowest common ancestor of the two senses s_1, s_2 :

$$\text{S\_Low\_Com\_Anc}(s_1,s_2) \in \text{A} \;\text{ and}$$

$$\text{depth}\,(\text{S\_Low\_Com\_Anc}(s_1, s_2)) \geq \underset{s \in \text{A}}{\text{Max}}\,[\text{depth}\,(s)].$$

We denote by S(w) the set of senses in WordNet of the word w and the LCA of two words w_1 and w_2 as W_Low_Com_Anc (w_1,w_2). W_Low_Com_Anc(w_1,w_2) is an S_Low_Com_Anc(s_i, s_j) / $\forall$ $s_1 \in$ S(w_1) and $s_2 \in$ S(w_2) : depth (S_Low_Com_Anc(s_i, s_j)) $\geq$ Max [depth (S_Low_Com_Anc(s_1, s_2))] (s_1 ranges over S(w_1) and s_2 ranges over S(w_2)).

Informally, the method consists of: (a) finding LCA by means of hypernym links and (b) going down by means of hyponym links in order to keep the appropriate sense.

Our main goal was to test such a method on a large scale by processing the Ohsumed corpus and evaluating the accuracy of results. This does not prevent us from proposing additional solution for the cases where this method records failures. The Analysis of the overall problem gives rise to three major problems, as follows:

Case1: **Single LCA.** Only one LCA is found and it is an ancestor for only one sense for each one of the two words in the pair. This is the most frequent case.

Solution: Once the LCA is identified, go down by hyponymy relationships and keep the senses of which the LCA is the ancestor.

Case 2: **Multiple LCA's.** In practice it is not trivial to find one LCA at each time, many cases include pairs of words with multiple LCA's. For example, the pair (*male*, *female*) has two LCA's: *animal* and *person*. *Animal* is the ancestor of the first sense (synset) of *male* and the first sense (synset) of *female*, while *person* is an ancestor of the second sense of *male* and the second sense of *female*. Which ancestor should we choose?

Solution: Compute the frequencies of occurrence of each one of the multiple LCA's and its synonyms in the corpus i.e. Ohsumed corpus. In other words, we compute the frequencies of occurrence in the Ohsumed corpus of the elements belonging to the synset representing the LCA. The LCA whose synset has the high synonyms frequency will be chosen.

Case 3: **Multiple senses.** As with case 1, we deal with only one LCA, but at least one of the two words has several senses which have this LCA as an ancestor. For example, the pair (*plasma*, *fluid*) has as LCA *substance*. It is the most specific common ancestor of all the following four combinations:1st sense of *plasma* and 1st sense of *fluid*, 1st sense of *plasma* and 2nd sense of *fluid*, 2nd sense of *plasma* and 1st sense of *fluid*, 2nd sense of *plasma* and 2nd sense of *fluid*. Which of these senses should be chosen?

Solution: we adopt the same idea as in Case 2. Compute for each one of these senses the frequencies of occurrence in the Ohsumed corpus of the element belonging to the corresponding synset. The synset with the highest value is chosen.

The underlying claim of the method of computing synonyms is as follows. Each sense in WordNet is represented by a set of synonyms (synset), a gloss and eventual examples. In front of multiple choices of senses, we select the sense which is the most often used in medicine. Finding the most used synset in the Ohsumed corpus performs this task. In other words, the appropriate synset is that one whose synonyms are the most used in Ohsumed, and consequently the associated sense is assumed to be the most prevalent in the medical domain.

5. IMPLEMENTATION

The architecture of the adopted approach is described in figure 3. In stage 1, we start by processing the Ohsumed corpus in order to extract all conjunctions, from which we constitute a set of word pairs (the two nouns of the conjunction). The extracted word pairs include general and medical terms. For a given word, to decide if it is a medical or a general term, we compare the number of its occurrences in Ohsumed against the corresponding frequency in BNC (British National Corpus). The words which occur more in Ohsumed than in the BNC are considered to be more domain-specific. These two tasks are performed respectively by the two modules: "Noun_CJC_Noun extraction module" and "Terms frequencies extraction module". Note that this stage only extracts word forms and does not make any judgment about which senses of these words are important.

The most important part is stage 2. In section 4, we explained how we dealt with the three cases of the above section, and how the different result sets were constructed. The "similarity module" finds the LCA for each word pair and outputs two sets:

$$\text{Success\_Set} = \{(w_1, w_2) /\ w_{1\_\text{CJC}\_}w_2 \in \text{ohsumed}$$
$$\text{and } \{ \text{W\_Low\_Com\_Anc}(w_1, w_2) \} \neq \varnothing \ \}.$$

$$\text{Failure\_Set} = \{(w_1, w_2) /\ w_{1\_\text{CJC}\_}w_2 \in \text{ohsumed}$$
$$\text{and } \{ \text{W\_Low\_Com\_Anc}(w_1, w_2) \ \}=\varnothing \}.$$

The Success_Set contains all word pairs for which at least one LCA has been identified. The Failure_Set contains all word pairs for which no common ancestor has been found.

From this set we extract a set of completely new words, i.e. those terms without entries in WordNet. (Mapping these terms into WordNet correctly is planned as a topic for future work.) We have decided to treat the cases cited above as follows: case 1 will be treated alone while case 2 and case 3 will be treated together. Consequently the Success_Set is divided in two subsets:

The first one i.e. One_L_C_A_No_Multi_Senses contains the elements of case 1 cited in the above section and is defined in Table 1.

Multi_L_C_A_&&_Multi_Sense is the second one that contains both the pairs representing case 2 (multiple ancestors) and case 3 (multiple senses); defined in Table 1.

In stage 3, we focus on the set of successful results. The "right medical sense checking module" is made up of two modules, evaluated by an expert of the domain.

The first module focuses on the elements of the One_L_C_A_No_Multi_Senses set and the other on the Multi_L_C_A_&&_Multi_Senses set. Each one of these two modules records the senses from WordNet chosen by our methods and allows the domain expert to answer the following question: "For which and how many words with more than one possible sense in WordNet have we successfully chosen to keep the appropriate sense for medicine?".

The module dealing with the elements of the Multi_L_C_A_&&_Multi_Senses set performs the technique of computing synonyms presented in the above section.

Table 1. The two subsets of Success_Set.

The set	The definition
One_L_C_A_ No_Multi_Senses	$\{(w_1, w_2)/(w_1, w_2) \in$ Success_Set and $\vert \{$W_Low_Com_Anc$(w_1, w_2)\} \vert =1$, and $\exists!\ s_i \in S(w_1)$, $\exists!\ s_j \in S(w_2)$ / W_Low_Com_Anc(w_1,w_2) = S_Low_Com_Anc$(s_i,s_j)\}$.
Multi_L_C_A_&&_Multi_Senses	$\{(w_1, w_2)/(w_1, w_2) \in$ Success_Set and $\vert \{$W_Low_Com_Anc$(w_1, w_2)\} \vert >1 \}$ $\cup$ $\{(w_1, w_2) / (w_1, w_2) \in$ Success_Set and $\vert \{$ W_Low_Com_Anc $(w_1, w_2)\} \vert =1$ and $(\exists\ S'(w_1, w_2) \subseteq (S(w_1)xS(w_2))$ / $\forall\ (s_i, s_j) \in S'(w_1, w_2)$, W_Low_Com_Anc$(w_1,w_2)$ = S_Low_Com_Anc(s_i,s_j) and $\vert\{ S'(w_1, w_2) \}\vert >1$) }.

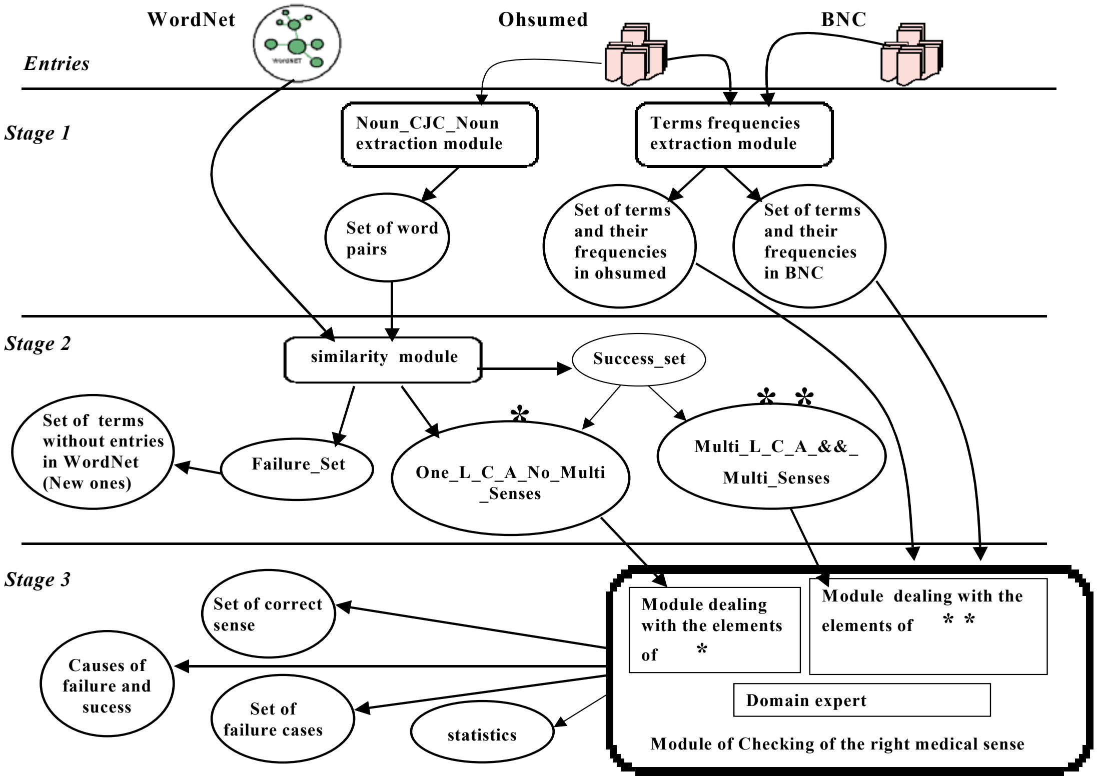
WordNet
Ohsumed
BNC
Entries
Stage 1
Noun_CJC_Noun extraction module
Terms frequencies extraction module
Set of word pairs
Set of terms and their frequencies in ohsumed
Set of terms and their frequencies in BNC
Stage 2
similarity module
Success_set
Set of terms without entries in WordNet (New ones)
Failure_Set
*
One_L_C_A_No_Multi _Senses
* *
Multi_L_C_A_&&_ Multi_Senses
Stage 3
Set of correct sense
Causes of failure and success
Set of failure cases
statistics
Module dealing with the elements of *
Module dealing with the elements of * *
Domain expert
Module of Checking of the right medical sense

Figure 3. The architecture of the adopted approach.

6. RESULTS AND ANALYSIS

From the Ohsumed Corpus [7] we extracted 57887 word pairs that were pairs of nouns separated by a conjunction, giving a total of 115774 words.

In stage 2, the cardinality of the Success_Set was 34220 pairs, and that of the Failure_Set was 23667 pairs. The failure of finding Common Ancestor of Failure_Set elements could be attributed to one of two causes.

Firstly, the words of many pairs like: *(tumor, tissue), (fertilization, gamete),* (*rhinitis, asthma*), (*infection, syndrome*), (*infarction, stroke*), are closely related in a medical domain, but are structurally very distant and have no hierarchical links between them. This is sometimes described as "The Tennis Problem". "The tennis problem" was noted by [4] as a phenomenon occurring in WordNet where related words could occur in two completely different parts of the ontology with no apparent link between them, e.g. *ball boy* could occur as a descendant of *male child* and *tennis ball* as a descendant of *game equipment*, so on a purely hierarchical basis, these words are deemed to be dissimilar despite an obvious semantic relation. In other words, WordNet does not offer the possibility of associating them all as concepts related to tennis. Stevenson [12] proposes a solution of this problem by adding thesaural relations to the noun taxonomy in WordNet. He proceeds by grouping noun synsets witch are related by topic or domain. Then he added new links between noun synsets belonging to the same group.

Secondly, many new words were discovered from Ohsumed. Many word pairs include new words, which aren't addressed by WordNet. We have found 8547 new words e.g. teicoplamin, thromboxane, prostacyclin, thymocyte, and vancomycin.

Now, we focus on the elements having at least one LCA, contained in the Success_Set . We have seen that this set is subdivided into two sets: One_L_C_A_No_Multi_Senses and Multi_L_C_A_&&_Multi_Senses. Note that our attention will be concentrated on the words with several senses in WordNet, in order to see the efficacy of the presented methods to get the appropriate sense for the medical domain. Table 2 shows for each one of the two previous sets the number of words with multiple senses and those with one sense. Table 2 shows clearly that the case1 category is the most frequent.

Table 2. The number of words with multiple senses and those with only one sense, for the two sets resulting from Success_Set.

	Words with more than one sense in WordNet	Words with only one sense in WordNet
One_L_C_A_ No_Multi_Senses: 20191 pairs ≡ 40382 words	26399 words	13983 words
Multi_L_C_A_&&_Multi_Senses : 14029 pairs ≡ 28058 words	3048 words	25010 words

In the following, we will analyse and comment on each set in turn. For each set we present the results and accuracies:

- Considering only the medical terms belonging to word pairs, and
- Considering only the general terms.

Due to the size of the Success_Set, a subset of 126 pairs was selected as a sample. This sample contains elements from the two sets.

If we refer to the results tables, among the 252 words, we found 188 medical terms and 64 more general terms. After obtaining results, for a given word we analyse the specificities and characteristics of the pair in which it belongs. Our interest is to know for the two words in the pair:

- Are they medical terms or general terms?
- What senses does WordNet associate to them?
- Are these only general senses, only medical senses, or both?

Result and Analysis of One_L_C_A_ No_Multi_Senses (the most frequent case)

Table 3 shows the results obtained for the 84 pairs of the sample and belonging to this set. It includes 136 domain-specific words and 32 general ones. For example, from the 136 specific words, 81 of them have multiple senses in WordNet; we get success in keeping the right sense for the medical domain in 59 cases (72%). Table 3 shows also results about words with one sense in WordNet, though a detailed analysis of these results is left to a subsequent project.

Table 3. Results of failure and success for One_L_C_A_ No_Multi_Senses set (the most frequent case). The results are presented for specific and general words, when considering words with multiple senses and those with only one sense.

	Words with more than one sense in WordNet		Words with only one sense in WordNet	
	Success	Failure	Success	Failure
Specific words: 136	59	22	54	01
General words: 32	11	16	03	02

Comments

The ACCURACY and RECALL values of words with multiple senses are respectively: 72% and 83% for specific terms; and 40% and 47% for general terms. For example, the word *stroke* has 10 senses in WordNet, our method has succeed to get the 3rd sense: "stroke, apoplexy, cerebrovascular accident, CVA -- (a sudden loss of consciousness resulting when the rupture or occlusion of a blood vessel leads to oxygen lack in the brain", which is the appropriate sense for medicine. We noticed that in most cases where we have two specific words and they are from the same part of the WordNet hierarchy (act, or state…), we have high chances of getting the right sense (if a medical sense exits). If we take the word *failure*, we were successful when it was associated with the domain-specific word *disease*, whereas when it was associated with the general word *death*, our method chose a sense completely foreign to medicine.

There are other cases, where the two words have medical senses in WordNet, but the method provides the wrong sense. Examples include the words *sign* and *symptom*, for them the sense meaning *indication* was chosen. The cause is that they have other senses for others domain which led to the discovery of a common ancestor which is deeper in the WordNet hierarchy than the ancestor from the medical domain. The most frequent cases of such failure are due to WordNet's omission of medical senses for many words, like the words *alpha* and *beta*.

All these comments were for the specific terms. Concerning the general terms, the accuracies are less than the specific ones. This is due to the general senses associated for most general words. But there are some exceptions, like the pair (*head*, *neck*). Although *head* has no fewer than 32 senses, the method has succeeded in getting a correct sense for *head* and also the right one for *neck* (*neck* has 4 senses).

Result and Analysis of Multi_L_C_A_&&_Multi_Senses

This set contains 42 pairs (84 words), out of the 126 that were selected for evaluation. Among these words, there are 52 medical terms and 32 general terms (Table 4). This set include the pairs which have multiple LCA ancestors and others which have only one LCA but with several senses with this LCA as ancestor. We noted that this later case is more frequent than the first one (pairs with multiple LCA).

Table 4. Results of failure and success for Multi_L_C_A_&&_Multi_Senses set. The results are presented for specific and general words, when considering words with multiple senses and those with only one sense.

	Words with more than one sense in WordNet		Words with only one sense in WordNet	
	Success	failure	success	failure
Specific words : 52	20	23	09	00
General words : 32	13	18	01	00

Comments

The ACCURACY and RECALL values of words with multiple senses are respectively: 46% and 51% for specific terms; and 41% and 52% for general terms. In many cases, we were in a situation where the word has a medical sense, but the method selects the wrong one from WordNet. After analysis, we realised that the cause was not the base of the synonym idea, but the cause in many cases is due to the chosen sense which has a synset with many synonyms in comparison with the medical sense. In other words, WordNet doesn't include for each synset the same number of synonyms. For example: the pair (*diagnosis*, *treatment*) has as LCA the concept *act*. The 1st, 2nd, and 4th senses of *treatment* all have *act* as an ancestor. After applying the method of computing frequencies of synset elements, the chosen one was the 4th, because the medical sense is the 1st one with only one word (*treatment*) in its synset, while the 4th sense has 3 synonyms in its synset. In other cases, simply because WordNet doesn't include any medical sense for a given word, all senses are general, so the chosen one can be any sense.

We quote other failures: The case where, among the candidate senses, the medical sense was not included. The cause has as origin in the method of finding the lowest common ancestor. In many cases the origin is the tennis problem, which doesn't allow the method to include the medical sense as a candidate sense for the method of computing synonyms. For example: for the pair (*relaxation*, *contraction*), the candidates senses for *relaxation* are: the 4th and 7th ones, while the 1st sense that is for physiology and which can be very well adopted to medicine, was not chosen, because it is a descendant from the *phenomenon* part of the hierarchy and *contraction* doesn't have any sense in this hierarchy.

In spite of these errors, we have still recorded good results (but not as many as those recorded for the elements of One_L_C_A_No_Multi_Senses set) where our method selected the appropriate medical senses. For example, in the case of the pair (*plasma*, *fluid*), the method has chosen the right sense for the two words and has avoided a variety of terms that are not suitable ones for medicine.

7. CONCLUSION

One of the most important object of study in computer science is the "how to give sense to objects" also known as the "semantic problem". Most of the actual research in several domains e.g. Artificial Intelligence, Natural Language Processing, Data Mining, Data Ware House, Semantic Web, Web Mining etc gravitate towards the semantic problem. Ontology i.e. the science of the study of the Being, offers some solutions to the semantic problem. An ontology provides a consensual concepts specifications for a specific domain shared by a group of people. WordNet is a lexical ontology i.e. a reference system that identifies words senses. WordNet is a general-purpose system. In this paper we study the possibility of mapping WordNet onto a specific domain i.e. medical domain. In semantic terms the problem is: "how to distinguish a domain specific sense from a general purpose sense". The presented method extracts phrases from the Ohsumed corpus that match the "Noun_CJC_Noun" patterns, and exploits the internal architecture of WordNet with the LCA relation to get the right sense. For the One LCA and No Multi Senses case the ACCURACY and RECALL are respectively: 72% and 83% for specific terms and those for general ones are 40% and 47%. For the Multi LCA and Multi Senses case the ACCURACY and RECALL values of words with multiple senses are respectively: 46% and 51% for specific terms and those for general ones are 41% and 52%. These results are positive and point out that the adopted solution is effective. Failure results are due to general and well-known semantic problems such as the tennis problem. Because of the handicraft of the analysis, future works will concern systematic and case-by-case analysis of the overall Success_Set. This will help us to characterize successes and failures, and to improve performance.

ACKNOWLEDGMENTS

The authors would like to thank Beate Dorow from Stuttgart University for providing the word pairs corresponding to the Noun_CJC_Noun extracted from Ohsumed corpus.

REFERENCES

[1] A. Budanitsky, and G. Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In Workshop on WordNet and Other Lexical Resources, 2nd of the North American Chapter of the ACL, Pittsburgh, June 2001

[2] P. Buitelaar and B. Sacaleanu: Extending Synsets with Medical Terms. In Proceedings of the First International Conference on Global WordNet, Mysore, India, 2002.

[3] Diab, Mona. The Feasibility of Bootstrapping an Arabic WordNet leveraging Parallel Corpora and an English WordNet. Proceedings of the Arabic Language Technologies and Resources, NEMLAR, Cairo 2004.

[4] C. Fellbaum, editor. WordNet: An electronic lexical database. MIT Press, Cambridge, MA, 1998.

[5] J. Hartmann, P. Spyns, A Giboin, D. Maynard, R. Cuel, M. Carmen Suarez de Figueroa and Y. Sure. Methods for Ontology Evaluation, KnowledgeWeb Deliverable #D1.2.3, Karlsruhe, 2004.

[6] Hearst, M. and H. Schütze. Customizing a lexicon to better suit a computational task. In Proceedings of the Special Interest Group on the Lexicon, Association for Computational Linguistics (ACL-SIGLEX Workshop). Columbus, OH, 1993.

[7] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In Proceedings of the 17th annual conference on Research and Development in Information Retrieval (SIGIR-94), pages 192–201, 1994.

[8] A. Lehireche et N.Doumi, Une ontologie pour le lexique arabe, in proceeding du 2ème congrès international de "l'ingénierie de la langue arabe et de l'ingénierie de la langue", CRSTDLA, Université d'Alger, juin 2005.

[9] G. A. Miller and W. G. Charles. Contextual correlates of semantic similarity. Language and cognitive Processes 1991, 6(1):1-28.

[10] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::similarity -measuring the relatedness of concepts. In Proc. of Fifth Annual Meeting of the North American Chapter of the ACL (NAACL-04), Boston, MA, 2004.

[11] P. Resnik, Semantic Similarity in a Taxonomy: An Information-Based Measure and its application to Problems of Ambiguity in Natural Language. Journal of artificial intelligence research, 11:93-130, 1999.

[12] M. Stevenson. Augmenting Noun Taxonomies by Combining Lexical Similarity Metrics. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING-02), Taipei, Taiwan, 2002

[13] P. Vossen. Introduction to EuroWordNet. Computers and the Humanities, 32(2-3):73–89, 1998.

[14] D. Widdows and B. Dorow. A Graph Model for Unsupervised Lexical Acquisition. Appeared in 19th International Conference on Computational Linguistics (*COLING* 19), Taipei, Taiwan, 2002.

[15] D. Widdows, Geometry and Meaning, CSLI publications, 2004.