Automatic Adaptation of WordNet to Sublanguages and to Computational Tasks

Roberto Basili(+) Alessandro Cucchiarelli (*) Carlo Consoli (+)

Maria Teresa Pazienza (+) Paola Velardi (#)

(+) Universita' di Roma Tor Vergata (ITALY)
(*) Universita' di Ancona (ITALY)
(#) Universita' di Roma, La Sapienza (ITALY)

Abstract

Semantically tagging a corpus is useful for many intermediate NLP tasks such as: acquisition of word argument structures in sublanguages, acquisition of syntactic disambiguation cues, terminology learning, etc. Semantic categories allow the generalization of observed word patterns, and facilitate the discovery of recurrent sublanguage phenomena and selectional rules of various types. Yet, as opposed to POS tags in morphology, there is no consensus in literature about the type and granularity of the category inventory. In addition, most available on-line taxonomies, as WordNet, are over ambiguous and, at the same time, may not include many domain-dependent senses of words. In this paper we describe a method to adapt a general purpose taxonomy to an application sublanguage: first, we prune branches of the Wordnet hierarchy that are too "fine grained" for the domain; then, a statistical model of classes is built from corpus contexts to sort the different classifications or assign a classification to known and unknown words, respectively.

1 Introduction

Lexical learning methods based on the use of semantic categories are faced with the problem of overambiguity and entangled structures of Thesaura and dictionaries. WordNet and Roget's Thesaura were not initially conceived, despite their success among researchers in lexical statistics, as tools for automatic language processing. The purpose was rather to provide the linguists with a very refined, general purpose, linguistically motivated source of taxonomic knowledge. As a consequence, in most on-line Thesaura words are extremely ambiguous, with very subtle distinctions among senses. High ambiguity, entangled nodes, and asymmetry have already been emphasized in (Hearst and Shutze, 1993) as being an obstacle to the effective use of on-line Thesaura in corpus linguistics. In most cases, the noise introduced by overambiguity almost overrides the positive effect of semantic clustering. For example, in (Brill and Resnik, 1994) clustering PP heads according to WordNet synsets produced only a 1% improvement in a PP disambiguation task, with respect to the non-clustered method. A subsequent paper (Resnik, 1997) reports of a 40% precision in a sense disambiguation task, always based on generalization through WordNet synsets. Context-based sense disambiguation becomes a prohibitive task on a wide-scale basis, because when words in the context of anambiguous word are replaced by their synsets, there is a multiplication of possible contexts, rather than a generalization. In (Agirre and Rigau, 1996) a method called Conceptual Distance is proposed to reduce this problem, but the reported performance in disambiguation still does not reach 50%. On the other hand, (Dolan, 1994) and (Krovetz and Croft, 1992) claim that fine-grained semantic distinctions are unlikely to be of practical value for many applications. Our experience supports this claim: often, what matters is to be able to distinguish among *contrastive* (Pustejowsky, 1995) ambiguities of the *bank_river bank_organisation* flavor. The problem however is that the notion of "contrastive" is domain-dependent. Depending upon the sublanguage (e.g. medicine, finance, computers, etc.) and upon the specific NLP application (e.g. Information Extraction, Dialogue etc.) a given semantic label may be too general or too specific for the task at hand. For example, the word *line* has 27 senses in WordNet, many of which draw subtle distinctions e.g. *line of work* (sense 26) and *line of products* (sense 19). In an

application aimed at extracting information on new products in an economic domain, we would be interested in identifying occurrences of such senses, but perhaps all the other senses could be clustered in one or two categories, for example *Artifact*, grouping senses such as: *telephone-line*, *railway* and *cable*, and *Abstraction*, grouping senses such as *series*, *conformity* and *indication*. Vice versa, if the sublanguage is technical handbooks in computer science, we would like to distinguish the *cable* and the *string* of words senses (7 and 5, respectively), while any other distinction may not have any practical interest.

The research described in this paper is aimed at providing some principled, and algorithmic, methods *to tune a general purpose taxonomy to specific sublanguages and domains.*

In this paper, we propose a method by which we select a set of core semantic nodes in the WordNet taxonomy that "optimally" describe the semantics of a sublanguage, according to a scoring function defined as a linear combination of general and corpus-dependent performance factors. The selected categories are used to prune WordNet branches that appear, according to our scoring function, less pertinent to the given sublanguage, thus reducing the initial ambiguity. Then, we learn from the application corpus a statistical model of the core categories and use this model to further tune the initial taxonomy. Tuning implies two actions:

- The first is to attempt a reclassification of relevant words in the corpus that are not covered by the selected categories, i.e., words belonging exclusively to pruned branches. Often, these words have domain-dependent senses that are not captured in the initial WordNet classification (e.g. the software sense of release in a software handbooks sublanguage). The decision to assign an unclassified word to one of the selected categories is based on a strong detected similarity between the contexts in which the word occurs, and the statistical model of the core categories.

- The second is to further reduce the ambiguity of words that still have a high ambiguity, with respect to the other words in the corpus. For example, the word *stock* in a financial domain still preserved the gunstock sense, because instrumentality was one of the selected core categories for the domain. The expectation of this sense may be lowered, as before, by comparing the typical contexts of *stock* with the acquired model of instrumentality.

In the next sections, we first describe the algorithm for selecting core categories. Then, we describe the method for redistributing relevant words among the nodes of the pruned hierarchy. Finally, we discuss an evaluation experiment.

2 Selection of core categories from WordNet

The first step of our method is to select from WordNet an inventory of *core categories* that appear particularly appropriate for the domain, and prune all the hierarchy branches that does not belong to such core categories. This choice is performed as follows:

Creation of alternative sets of balanced categories

First, an iterative method is used to create alternative sets of balanced categories, using information on words and word frequencies in the application corpus. Sets of categories have an increasing level of generality. The set-generation algorithm is an iterative application of the algorithm proposed in (Hearst and Shutze, 1993) for creating WordNet categories of a fixed average size. In short[1], the algorithm works as follows: Let C be a set of WordNet synsets s, W the set of different words (nouns) in the corpus, $P(C)$ the number of words in W that are instances of C, weighted by their frequency in the corpus, UB and LB the upper and lower bound for $P(C)$. At each iteration step i, a new synset s is added to the current category set C_i, iff the weight of s lies within the current boundaries, that is, $P(s) \leq UB_i$ and $P(s) \geq LB_i$. If $P(s) \geq UB_i$ s is replaced in C_i by its descendants, for which the same constraints are verified. If $P(s) \leq LB_i$, s is added to a list of "small" categories $SCT(C_i)$. In fact, when replacing an overpopulated category by its sons, it may well be the case that some of its sons are under populated.

[1] The procedure new_cat(S) is almost the same as in (Hearst and Shutze, 1993). For sake of brevity, the algorithm is not explained in much details here.

Scoring Alternative Sets of Categories

Second, a scoring function is applied to alternative sets to identify the core set. The core set is modeled as the linear function of four performance factors: *generality*, *coverage of the domain*, *average ambiguity*, and *discrimination power*. For a formal definition of these four measures, see (Cucchiarelli and Velardi, 1997). We provide here an intuitive description of these factors:

Generality (G): In principle, we would like to represent the semantics of the domain using the highest possible level of generalization. A small number of categories allows a compact representation of the semantic knowledge base, and renders word sense disambiguation more simple. On the other side, over general categories fail to capture important distinctions. The Generality is a gaussian measure that mediates between over generality and overambiguity.

Coverage (CO) This is a measure of the coverage that a given category set C_i has over the words in the corpus. The algorithm for balanced category selection does not allow a full coverage of the words in the domain: given a selected pair $< UB, LB >$, it may well be the case that several words are not assigned to any category, because when branching from an overpopulated category to its descendants, some of the descendants may be under populated. Each iterative step that creates a C_i also creates a set of under populated categories $SCT(C_i)$. Clearly, a "good" selection of C_i is one that minimizes this problem (and has therefore a "high" coverage).

Discrimination Power (DP): A certain selection of categories may not allow a full discrimination of the lowest-level senses for a word (leaves-synsets hereafter). For example, if psychological_feature is one of the core categories, and if we choose to tag a corpus only with core categories, it would be impossible to discriminate between the *business-target* and *business-concern* senses. Though nothing can be said about the practical importance of discriminating between such two synsets, in general a good choice of C_i is one that allows as much as possible the discrimination between low level senses of ambiguous words.

Average Ambiguity (A) : Each choice of C_i in general reduces the initial ambiguity of the corpus. In part, because there are *leaves-synsets* that converge into a single category of the set, in part because there are leaves-synsets of a word that do not reach any of these categories. Though in general we don't know if, by cutting out a node, we are removing a set of senses interesting (or not) for the domain, still in principle a good choice of categories is one that reduces as much as possible the initial ambiguity. The cumulative scoring function for a set of categories C_i is defined as the linear combination of the performance parameters described above:

$$Score(C_i) = \alpha G(C_i) + \beta CO(C_i) + \gamma DP(C_i) + \delta \frac{1}{A(C_i)} \quad (1)$$

Estimation of model parameters and refinements

An interpolation method is adopted to estimate the parameters of the model against a reference, correctly tagged, corpus (SemCor, the WordNet semantic concordance). The performance of alternative inventories of core categories is evaluated in terms of *effective reduction of overambiguity*. This measure is a combination of the system precision at pruning out spurious (for the domain) senses, and the global reduction of ambiguity. Notice that we are not measuring the precision of sense disambiguation in contexts, but simply the precision at reducing a-priori the set of possible senses for a word, in a given domain.

The method above is weakly supervised: the parameters estimated have been used without re-estimation to capture core categories in other domains such as Natural Science and a UNIX manual. Details on portability of this choice are in (Cucchiarelli and Velardi, forthcoming 1998).

In the different experiments, the best performing choice of core categories is the one with an upper population of 62.000 words (frequency weighted). This corresponds to the following list of 14 categories:

```
num_cat=14 i=61 UB=62000 LB=24800 N=2000 k=1000 h=0.40

person, individual, someone, mortal, human, soul
instrumentality, instrumentation
attribute
written_communication, written_language
message, content, subject_matter, substance
measure, quantity, amount, quantum
action
activity
group_action
organization
psychological_feature
possession
state
location
```

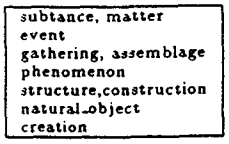
This selection of core categories is measured to have the following performance:

Precision: 77.6%

Reduction of Ambiguity: 37%

Coverage: 78%

In (Cucchiarelli and Velardi, forthcoming 1998) a method is proposed to automatically increase the coverage of the core set with an additional set of categories, selected from the set of under populated categories $SCT(C_i)$ (see step 1 of the algorithm). With the extension:



the following performance is obtained:
Precision: 78,9%
Reduction of Ambiguity: 26%
Coverage: 93%

With some manual refinement of the extended set , the precision rises to over 80%. Obtaining a higher precision is difficult because, neither SemCor nor WordNet can be considered a golden standard. In a recent workshop on semantic texts tagging (TAGWS 1997), the difficulty of providing comprehensible guidelines for semantic annotators in order to avoid disagreement and inconsistencies was highlighted. On the other side, there are many redundancies and some inconsistencies in WordNet that makes the task of (manual) classification very complex. To make an example, one of the detected classification errors in our Wall Street Journal experiment was the selection of two possible core senses for the word *market*: : *organization* and *activity*. Vice versa, in the economic fragment of SemCor, *market* is consistently classifies as socio-economic-class, which happens not to be a descendent of any of these two categories. Our intuition when observing the specific examples was more in agreement with the automatic classification than with SemCor. Our feeling was that the selected core categories could, in many cases, represent a good model of classification for words that remained unclassified with respect to the "not pruned" WordNet, or appeared misclassified in our evaluation experiment.

In the next section we describe an method to verify this hypothesis and, at the same time, to further tune WordNet to a domain.

3 Redistribution of words among core categories

The purpose of the method described hereafter is twofold:

- The first is to attempt a reclassification of words that are not classified, or appeared as misclassified, with respect to the "original" WordNet.
- The second is to further reduce the ambiguity of words that are still very ambiguous with respect to the "pruned" WordNet. The general idea is that ambiguity of words is reduced in a specific domain, and enumeration of all their senses is unnecessary. Second, some words function as sense primers for others. Third, raw contexts of words provide a significant bundle of information to guide disambiguation.

To verify this hypothesis systematically we need to acquire from the corpus a contextual model of the core categories, and then verify to what extent certain "interesting" words (for example, unclassified words) adhere to the contextual model of one of such categories.

Our method, inspired by (Yarowsky, 1992), works as follows (see (Basili et al, 1997) for details) :

- Step 1. Select the most typical words in each core category:
- Step 2. Acquire the collective contexts of these words and use them as a (distributional) description of each category:
- Step 3. Use the distributional descriptions to evaluate the (corpus-dependent) membership of each word to the different categories.

Step 1 is carried out detecting the more significant (and less ambiguous) words in any of the core classes : these sets are called the kernel of the corresponding class. Rather than training the classifier on all the nouns in the learning corpus as in (Yarowsky, 1992), we select only a subset of *prototypical* words for each category. We call these words w the *salient words* of a

category C. We define the *typicality* $T_w(C)$ of w in C, as:

$$T_w(C) = \frac{N_{w,C}}{N_w} \quad (2)$$

where:
N_w is the total number of synsets of a word w, i.e. all the WordNet synonymy sets including w.
$N_{w,C}$ is the number of synsets of w that belong to the semantic category C, i.e. synsets indexed with C in WordNet.

The *typicality* depends only on WordNet. A *typical noun* for a category C is one that is either non ambiguously assigned to C in WordNet, or that has most of its senses (synsets) in C.

The *synonymy* S_w of w in C, i.e. the degree of synonymy showed by words other than w in the synsets of the class C in which w appears, is modeled by the following ratio:

$$S_w(C) = \frac{O_{w,C}}{O_w} \quad (3)$$

where:
O_w is the number of words in the corpus that appear in at least one of the synsets of w.
$O_{w,C}$ is the number of words in the corpus appearing in at least one of the synsets of w, that belong to C.

The *synonymy* depends both on WordNet and on the corpus. A noun with a high degree of synonymy in C is one with a high number of synonyms in the corpus, with reference to a specific sense (synset) belonging to C. *Salient nouns* for C are frequent, typical, and with a high synonymy in C. The *salient words* w, for a semantic category C, are thus identified maximizing the following function, that we call *Score*:

$$Score_w(C) = OA_w \times T_w(C) \times S_w(C) \quad (4)$$

where OA_w are the absolute occurrences of w in the corpus. The value of *Score* depends both on the corpus and on WordNet. OA_w depends obviously on the corpus.
The *kernel* of a category *kernel*(C), is the set of salient words w with a "high" $Score_w(C)$. In Table 1 some kernel words for the class *gathering,assemblage* are reported.

Step 2 uses the kernel words to build (as in (Yarowsky,1992)) a probabilistic model of a class: this model is based on the distribution of class relevance of the surrounding terms in typical contexts.

Table 1: Some kernel elements for class 17: gathering, assemblage

Score	Word	Score	Word
0.68835	executive	0.11108	business
0.55539	senate	0.11108	household
0.33828	public	0.10014	council
0.28485	court	0.08920	school
0.23815	family	0.08864	session
0.20869	commune	0.08780	form
0.14839	press	0.08667	town
0.11907	vote	0.07868	staff

In Step 3 a word is assigned to one, or more, classes according to the contexts in which it appears. Many contexts may enforce the selection of a given class, or multiple classifications are possible when different contexts suggest independent classes. For a given word w, and for each category C, we evaluate the following function, that we call *Domain Sense* (*DSense*(w, C)):

$$DSense(w, C) = \frac{1}{N} \sum_k Y(k, C) \quad (5)$$

where

$$Y(k, C) = \sum_{w' \in k} Pr(w', C) \times Pr(C) \quad (6)$$

where k's are the contexts of w, and w' is a generic word in k.
In (6), $Pr(C)$ is the (not uniform) probability of a class C, given by the ratio between the number of collective contexts for C [2] and the total number of collective contexts.

4 Discussion of the experiment

In this section we describe some preliminary results of an experiment conducted on the Wall Street Journal. We used 21 categories including 14 core categories plus 7 additional categories obtained with automatic extension of the best core set (see section 2). In experiment 1, we selected the 6 most frequent unclassified words in the corpus, and attempted a reclassification

[2] those collected around the kernel words of C

according to the contextual description of the 21 categories. In experiment 2, we selected the 6 most frequent and still very ambiguous (according to the pruned WordNet) words, and attempted a reduction of ambiguity. For each word w and each category C, we compute the $DSense(w, C)$ and then select only those senses that exhibit a membership value higher than the average membership of kernel words of C. The assignment of a word to a category is performed regardless of the current classification of w in the pruned WordNet.

The following Table 2 summarizes the results of experiment 1:

Table 2: Selected categories for some unclassified words

Word/freq	Selected categories
wall/447	gathering,written_communication,organization
pentagon/183	gathering,location,organization
people/973	gathering
airport/59	construction,location
congress/456	gathering,person

Table 3 reports on experiment 2. In column 3, selected categories are reported in decreasing order of class membership evidence.

In Table 2, notice the apparently "strange" classification of *wall*. The problem is that, in the current version of our system, proper nouns are not correctly detected (this problem will be fixed shortly) since in the Wall Street Journal there is no special syntactic tag for proper names. Erroneously, several proper names, such as *Wall Street*, *Wall Street Journal*, *Bush*, *Delta*, *Apple*, etc. were initially classified as common nouns, therefore causing some noise in the data that we need now to eliminate[3].

Table 3: Selected and Initial WN categories for some very ambiguous words

Word/freq	Initial WN categories	Selected categories
share/3473	written_communication possession group_action activity instrumentality	written_communication possession group_action - -
stock/3106	written_communication possession person natural_object instrumentality	written_communication possession person - -
price/2132	message,content possession attribute	message,content possession attribute
bank/1913	organization possession instrumentality natural_object	organization possession - -
business/1563	group_action organization gathering psych. feature activity	group_action organization gathering - -
bond/1366	- possession attribute phenomenon instrumentality	creation possession attribute - -

The word *wall* is in fact part of the complex nominals *Wall Street* and *Wall Street Journal*, and it is very interesting that, based on the context, the system classifies it correctly in the three categories: *gathering,written_communication,organization* Notice that the category: "*gathering, assemblage*" has somehow an unintuitive label, but in the WSJ domain this class includes rather uniform words, most of which refer to political organizations, as shown in Table 1.

In Table 3, it is shown that often some reduction of ambiguity is possible. However, some spurious senses survive, for example, the *progenitor* (person) sense of *stock*. It is very important that, in all the analyzed cases, the selected classes are a subset of the initial WordNet classes: remember that the assignment of a word to a category is performed only on the basis of its computed membership to that category. There is one example of additional detected sense (not included in the pruned WordNet), i.e. the sense *creation* for the word *bond*. Typical (for the domain) words in this class are: *plan*, *yeld*, *software*, *magazine*, *journal*, *issue*, etc. therefore, the creation sense seems appropriate.

Clearly, we need to perform a better (in the large) experimentation, but the first results seem encouraging. A large scale experiment requires, besides a better tuning of the statistical parameters and fixing some obvious bug (e.g. the identification of proper nouns), the preparation of a test set in which the correct classification of a large number of words is verified manually in the actual corpus contexts. Finally, experiments should be extended to domains other than WordNet. We already experimented the algorithm for core category selection on a UNIX corpus and on a small Natural Science corpus, but again, extending the complete experiment

[3] For example, the additional category *natural_object* was created because of the high frequency of spurious nouns as apple, delta, bush, etc.

to other corpora is not trivial for the required intensive linguistic and statistical corpus processing.

5 References

(Agirre and Rigau, 1996) E. Agirre and G. Rigau, Word Sense Disambiguation using Conceptual Density, proc. of COLING 1996

(Basili, Della Rocca, Pazienza, 1997) R. Basili, M.Della Rocca, M.T. Pazienza, Towards a Bootstrapping Framework for Corpus Semantic Tagging , in (TAGWS 1997)

(Basili et al. 1995b.) Basili R., M. Della Rocca, M.T. Pazienza, P. Velardi. "Contexts and categories: tuning a general purpose verb classification to sublanguages". Proceeding of RANLP95, Tzigov Chark, Bulgaria, 1995.

(Brill and Resnik, 1994) E. Brill and P. Resnik, A transformation-based approach to prepositional phrase attachment disambiguation, proc. of COLING 1994

(Chen and Chen, 1996) K. Chen and C. Chen, A rule-based and MT-oriented Approach to Prepositional Phrase Attachment, proc. of COLING 1996

(Cucchiarelli and Velardi, forthcoming 1998) Cucchiarelli A., Velardi P. "Finding a Domain-Appropriate Sense Inventory for Semantically Tagging a Corpus" Int. Journal of Natural Language Engineering, in press, 1998

(Cucchiarelli and Velardi, 1997) Cucchiarelli A., Velardi P. "Automatic Selection of Class Labels from a Thesaurus for an Effective Semantic Tagging of Corpora", 6th Conf. on Applied Natural Language Processing, ANLP97, Washington, April 1-3 1997

(Dolan, 1994) W. Dolan, Word Sense Ambiguation: Clustering Related Senses, Proc. of Coling 1994

(Fellbaum, 1997) C. Fellbaum, "Analysis of a hand-tagging task" in (TAGWS 1997).

(Hearst and Schuetze, 1993) M. Hearst and H. Schuetze, Customizing a Lexicon to Better Suite a Computational Task, ACL SIGLEX, Workshop on Lexical Acquisition from Text, Columbus, Ohio, USA, 1993.

(Yarowsky, D. 1992), "Word-Sense disambiguation using statistical models of Roget's categories trained on large corpora". Nantes: Proceedings of COLING 92.

(Krovetz and Croft, 1992) R. Krovetz and B. Croft, Lexical Ambiguity and Information Retrieval, in ACM trans. on Information Systems, 10:2, 1992

(Gale et al. 1992) Gale, W. K. Church and D. Yarowsky, One sense per discourse, in proc. of the DARPA speech and and Natural Language workshop, Harriman, NY, February 1992

(Pustejovsky, 1995) J. Pustejovsky, The generative Lexicon, MIT Press, 1995

(Resnik, 1995) P. Resnik, Disambiguating Noun Groupings with respect to Wordnet Senses, proc. of 3rd Workshop on Very Large Corpora, 1995

(Resnik, 1997) P. Resnik, Selectional reference and Sense disambiguation, in TAGWS97

(TAGWS 1997) Proceedings of the workshop "Tagging Text with Lexical Semantics: Why, What, and How?", published by ACL, 4-5 April 1997, Whashington, USA