Automatic Selection of Class Labels from a Thesaurus for an Effective Semantic Tagging of Corpora.

Alessandro Cucchiarelli †
†Università di Ancona - alex@inform.unian.it

Paola Velardi ‡
‡Università di Roma 'La Sapienza' - velardi@dsi.uniroma1.it

Abstract

It is widely accepted that tagging text with semantic information would improve the quality of lexical learning in corpus-based NLP methods. However available on-line taxonomies are rather entangled and introduce an unnecessary level of ambiguity. The noise produced by the redundant number of tags often overrides the advantage of semantic tagging. In this paper we propose an automatic method to select from WordNet a subset of domain-appropriate categories that effectively reduce the overambiguity of WordNet, and help at identifying and categorise relevant language patterns in a more compact way. The method is evaluated against a manually tagged corpus, SEMCOR.

1 Introduction

It is well known that statistically-based approaches to lexical knowledge acquisition are faced with the problem of low counts. Many language patterns (from simple co-occurrences to more complex syntactic associations among words) occur very rarely, or are never encountered, in the learning corpus. Since rare patterns are the majority, the quality and coverage of lexical learning may result severely affected.

The obvious strategy to reduce this problem is to generalise word patterns according to some clustering techniques. In the literature, two generalisation strategies have been adopted:

Distributional approaches: Several papers adopt distributional techniques to identify clusters of words according to some defined measure of similarity. Among these, in (Grishman and Sterling, 1994) a method is proposed to cluster syntactic triples, while in (Pereira and Tishby 1992, 1993) , (Dagan et al., 1994) pure bigrams are analysed.

The most intuitive evaluation of the effectiveness of distributional approaches to the problem of word generalization is presented in (Grishman and Sterling, 1994). In this paper it is argued that distributional (called also smoothing) techniques introduce a certain degree of additional error, because co-occurrences may be erroneously conflated in a cluster, and some of the co-occurrences being generalized are themselves incorrect. In general, the effect is a higher recall at the price of a lower precision.

Another drawback of these methods is that, since clusters have only a numeric description, they are often hard to evaluate on a linguistic ground.

Semantic tagging: Another adopted solution is to generalise the observed word patterns by grouping patterns in which words have the same semantic tag. Semantic tags are assigned from on-line thesaura like WordNet (Basili et al, 1996) (Resnik, 1995) , Roget's categories (Yarowsky 1992) (Chen and Chen, 1996), the Japanese BGH (Utsuro et al, 1993), or assigned manually (Basili et al, 1992)[1].

The obvious advantage of semantic tags is that words are clustered according to an intuitive principle (they belong to the same concept) rather than to some probabilistic measure. Semantic tagging has been proven useful for learning and categorising interesting relations among words, and for systematic lexical learning in sublanguages, as shown in (Basili et al, 1996) and (Basili et al, 1996b).

On the other hand, semantic tagging has a serious drawback, which is not solely due to the limited availability of on-line resources, but rather to the entangled structure of thesaura. Wordnet and Roget's thesaura have not been 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.

(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* flavour.

High ambiguity, entangled nodes, and asymmetry have already been emphasised 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. There are reported cases in which the use of WordNet worsened the performance of an automatic indexing method. Even context-based sense disambiguation becomes a prohibitive task on a wide-scale basis, because when words in the context of an ambiguous word are replaced by

[1] Manually assigning semantic tags if of course rather time-consuming, however on-line thesaura are not available in many languages, like Italian.

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 do not reach 50%.

A possible alternative is to manually select a set of *high-level* tags from the thesaurus. This approach is adopted in (Chen and Chen, 1996) and in (Basili et al, 1996) where only a dozen categories are used. As discussed in the latter paper, high-level tags reduce the problem of overambiguity and allow the detection of more regular behaviours in the analysis of lexical patterns. On the other hand, high-level tags may be overgeneral, and the acquired lexical rules, while usually perform well in the task of selecting the correct word associations (for example in PP disambiguation, or sense interpretation), are less capable of filtering out the noise. Overgeneral categories may even fail to capture contrastive ambiguities of words.

So far the manual selection of an appropriate set of semantic tags has been a matter of personal intuitions, but we believe that this task should be performed in a more principled, and automatic, way.

In this paper, we present a method for the selection of the "best-set" of WordNet categories for an effective, domain-tailored, semantic tagging of a corpus. The purpose of the method is to automatically select:

- A domain-appropriate set of categories, that well represent the semantics of the domain.
- A "right" level of abstraction, so as to mediate at best between overambiguity and overgenerality.
- A balanced (for the domain) set of categories, i.e. words should be evenly distributed among categories.

The second feature is the most important, since as we remarked so far, assigning semantic characteristics to words is very useful in lexical learning tasks, but *overambiguity is the major obstacle to an effective use of thesaura in semantic tagging.*

In the following sections, we define a method for the automatic selection of the "best-set" of WordNet categories, for nouns given an application corpus.

First, an iterative method is used to create alternative sets of *balanced* categories. Sets have an increasing level of generality. Second, a scoring function is applied to alternative sets to identify the "best" set. The best set is modelled as the linear function of four performance factors: *generality, coverage of the domain, average ambiguity,* and *discrimination power.* An interpolation method is adopted to estimate the parameters of the model against a reference, correctly tagged, corpus (SEMCOR). The performance of the selected set of categories is evaluated in terms of *effective reduction of overambiguity.*

The described method only requires a medium-range (stemmed) application corpus and a thesaurus. The model parameters are tuned against a reference correctly tagged corpus, but this is not strictly necessary if correctly tagged corpora are not available.

2 Selection of Alternative Sets of Semantic Categories from WordNet

The first step of the method is generating alternative sets of WordNet categories. Alternative sets are selected according to the following principles:

- Balanced categories: words must be uniformly distributed among categories of a set;
- Increasing level of generality: alternative sets are selected by uniformly increasing the level of generality of the categories belonging to a set;
- Domain-appropriateness: selected categories in a set are those pointed by (an increasingly large number of) words of the application domain, weighted by their frequency in the corpus.

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 its modified version, the algorithm is as follows[2]:

Let S be a set of WordNet *synsets* s, W the set of different words (nouns) in the corpus, P(s) the number of words in W that are instances of s, weighted by their frequency, UB and LB the upper and lower bound for P(s), N, h and k constant values.

```
i=1
UB=N
LB=UB*h;
do
{  initialise S with the set of WordNet topmost;
   initialise the set of categories C_i with
   the empty set;
   new_cat(S);
   if i=1 or C_i≠C_i-1 then add C_i to the set of Cat.
   i=i+1;
   UB=UB+k;
   LB=UB*h;
}
until C_i is not an empty set;

where:

new_cat(S):
for any category s of S
{
  if s does not belong to C_i then
  {
    if P(s) <= UB and P(s) >= LB
      then put s in the set C_i
      else if P(s) > UB
        then
        {
        let S' be the set of direct descendents of s
        new_cat(S')
        }
        else add s to SCT(C_i)
  }
}
```

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

N, h and k are the initial parameters of the algorithm. We experimentally observed that only h (the ratio between lower and upper bound) significantly modifies the resulting sets of categories (C_i): we established that a good compromise is h=0.4. SCT(C_i) is the set of "smaller" WordNet categories with P(s)<LB that do not belong to the C_i set (see next section).

3 Scoring Alternative Sets of Categories

The algorithm of section 2 creates alternative sets of balanced and increasingly general categories C_i. We now need a scoring function to evaluate these alternatives.
The following performance factors have been selected to express the scoring function:
Generality: In principle, we would like to represent the semantics of the domain using the highest possible level of generalisation. We can express the generality G'(C_i) as 1/DM(C_i), being DM(C_i) the average distance between the categories of C_i and the WordNet topmost synsets. Due to the graph structure of WordNet, different paths may connect each element c_{ij} of C_i with different topmosts, therefore we compute DM(C_i) as:

$$DM(Ci) = \frac{1}{n} * \sum_{j=1}^{n} dm(cij)$$

where dm(c_{ij}) is the average distance of each c_{ij} from the topmosts. Figure 1 illustrates a possible sysnsets hierarchical in which, for C_i={c_{i1} c_{i2}}, being dm(c_{i1})=(4+3)/2 =3.5 and dm(c_{i2})=3, DM(C_i)=(3+3.5)/2=3.25

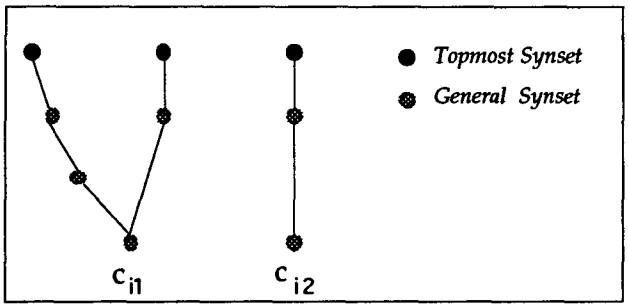

Figure 1 - An example of synsets hierarchy

As defined, G'(C_i) is a linear function (low values for low generality, high value for high generality), whilst our goal is to mediate at best between overspecificity and overgenerality. Therefore, we model the generality as G(C_i)=G'(C_i)*Gauss(G'(C_i)), where Gauss(G'(C_i)) is a Gauss distribution function computed by using the average and the variancy of G'(C_i) values over the set of all categories C_i, selected by the algorithm in section 2, normalised in the [0,1] interval.
Coverage: the algorithm of section 2, for any set C_i, does not allow a full coverage of the nouns 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 underpopulated. Each iterative step that creates a C_i also creates a set of underpopulated categories SCT(C_i). To ensure full coverage, these categories may be added to C_i, or alternatively, they can be replaced by their direct ancestors, but clearly a "good" selection of C_i is one that minimizes this problem. The coverage CO(C_i) is therefore defined as the ratio $N_c(C_i)/W$, where $N_c(C_i)$ is the number of words that reach at least one category of C_i
Discrimination Power: a certain selection of categories may not allow a full discrimination of the lowest-level senses for a word (*leaves-synsets* hereafter). Figure 2 illustrates an example. If C_i = {c_{i1} c_{i2} c_{i3} c_{i4}}, w_2 cannot be fully disambiguated by any sense selection algorithm, because two of its leaves-synsets belong to the same category c_{i2}. With respect to w_2, c_{i2} is overgeneral (though nothing can be said about the actual importance of discriminating between such two synsets).
We measure the discrimination power DP(C_i) as the ratio $(N_c(C_i)-N_{pc}(C_i))/N_c(C_i)$, where $N_c(C_i)$ is the number of words that reach at least one category of C_i, and $N_{pc}(C_i)$ is the number of words that have at least two leaves-synsets that reach the same category c_{ij} of C_i. For the example of figure 2 DP1, DP(C_i)=(3-1)/3=0.66.

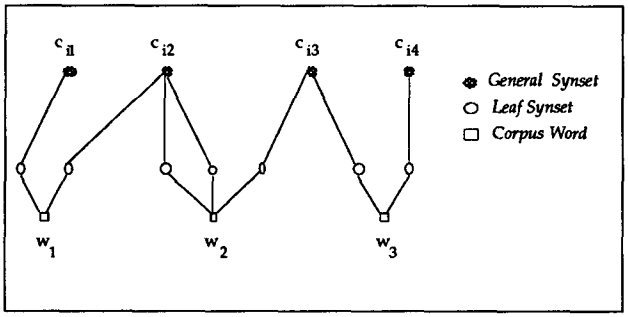

Figure 2 - An example of distribution of leaves-synsets among categories

Average Ambiguity: 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. This phenomenon is accounted for by the *inverse* of the average ambiguity A(C_i). The A(C_i) is measured as:

$$A(Ci) = \frac{1}{Nc(Ci)} * \sum_{j=1}^{Nc(Ci)} Cwj(Ci)$$

where $N_c(C_i)$ is the number of words that reach at least one category of C_i and, for each word w_j in this set, $Cw_j(C_i)$ is the number of categories of C_i reached.
In figure 2, the average ambiguity is 2 for the set C_i = {c_{i1} c_{i2} c_{i3} c_{i4}}, and is 5/3=1.66 for C_i = {c_{i1} c_{i2} c_{i3}}.

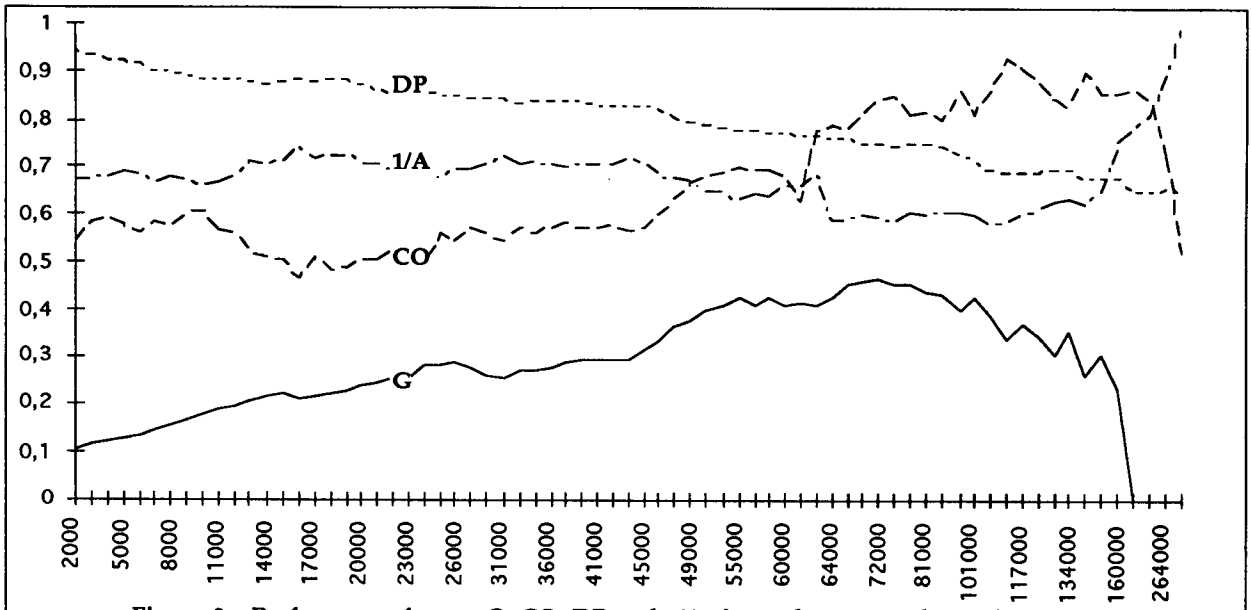

Figure 3 - Performance factors G, CO, DP and 1/A, for each generated set of categories

The *scoring function* for a set of categories C_i is defined as the linear combination of the performance parameters described above:

(1) $\text{Score}(C_i) = \alpha G(C_i) + \beta CO(C_i) + \chi DP(C_i) + \delta(1/A(C_i))$

Notice that we assigned a positive effect on the score (modelled by 1/A) to the ability of eliminating certain leaves-synsets and a negative effect (modelled by DP) to the inability of discriminating among certain other leaves-synsets. This is reasonable in general, because our aim is to control overgenerality while reducing overambiguity. However, nothing can be said on the appropriateness of a specific sense aggregation and/or sense elimination for a word. It may well be the case that merging two senses in a single category is a reasonable thing to do, if the senses do not draw interesting (for the domain) distinctions. Therefore eliminating a priori a sense of a word may be inappropriate in the domain.

The (1) is computed for all the generated sets of categories C_i, and then normalised in the [0,1] interval. The effectiveness of this model is estimated in the following section.

4 Evaluation Experiments and Discussion of the Data

The algorithm was applied to the 10,235 different nouns of the Wall Street Journal (hereafter WSJ) corpus that are classified in WordNet. Categories are generated with h= 0.4 and k=1,000. The cardinality of each set varies, but not uniformly, from 456 categories for UB=2000 (remember that words are frequency-weighted), to 1 category (i.e. the topmost *entity*) for UB=264,000. Medium-high level categories (those between 50,000 and 100,000 maximum words) range between 10-20 members for each set C_i.

Figure 3 plots the values of G, CO, DP and 1/A for the different sets of categories generated by the algorithm of Section 2. Alternative sets of categories are identified by their upperbound[3]. The figure shows that $DP(C_i)$ has a regular decreasing behaviour, while $1/A(C_i)$ is less regular. The coverage $CO(C_i)$ has a rather unstable behaviour due to the entangled structure of WordNet. We attempted slight changes in the definitions and computation of CO, DP and 1/A (for example, weighting words with their frequency), but globally, the behaviour remain as those in figure 3.

To compute the score of each set C_i, the parameters α,β,χ and δ in (1) must be estimated. To perform this task, we adopted a linear interpolation method, using SEMCOR (the semantically tagged Brown Corpus) as a reference corpus. In SEMCOR every word is unambiguously tagged with its leaf-synset.

To build a reference scoring function against which to evaluate our model parameters, we proceeded as follows:

- Since our categories are generated for an economic domain (WSJ) while SEMCOR is a tagged balanced corpus (the Brown Corpus), we extracted only the fragment of the corpus dealing with economic and financial texts. We obtained a reference corpus including 475 of the 1,235 nouns of the WSJ corpus.
- For each set of categories C_i generated by the algorithm in section 2, we computed on the reference corpus the following two performance figures:

Precision: For each C_i, let $W(C_i)$ be the set of words in the reference corpus covered by the set C_i. For each w_k in

[3]Remember that words are weighted by their frequency in the corpus. This seems reasonable, but in any case we observed that our results do not vary when counting each word only once.

$W(C_i)$, let $S(w_k)$ be the total set of leaves-synsets of w_k in WordNet, $SR(w_k)$ the subset of leaves-synsets of w_k found in the reference corpus, $SC(w_k)$ the subset of leaves-synsets that reach some of the categories of C_i. Let $WR(C_i) \subseteq W(C_i)$ be the set of w_k having $SC(w_k) \subset S(w_k)$. Following the algorithm:

```
for any w_k in WR(C_i)
{
   for any s_i in SR(w_k)
      {
      if s_i ∈ SC(w_k) then N+ = N+ + freq_i(w_k)
      N^tot = N^tot + freq(w_k)
      }
}
```

where $freq(w_k)$ is the number of occurrences of w_k in the reference corpus, the precision $Precision(C_i)$ is then defined as N^+ / N^{tot}. The precision measures the ability of each set C_i at correctly pruning out some of the senses of $W(C_i)$.

Global reduction of ambiguity: For each C_i, let $S(W_i)$ be the total number of WordNet leaves-synsets reached by the words in $WR(C_i)$, and $SC(W_i) \subseteq S(W_i)$ the set of these synsets that reach some category in C_i. By tagging the corpus with C_i, we obtain a reduction of ambiguity measured by:

$$GRAmb(C_i) = (card(S(W_i)) - card(SC(W_i))) / card(S(W_i))$$

where card (X) is the number of elements in the set X

Starting from these two performance figures, the global performance function $Perf(C_i)$ is measured by:

$$(2) \quad Perf(C_i) = Precision(C_i) + GRAmb(C_i)$$

The (2) is computed for all the generated sets of categories C_i, and then normalised in the [0,1] interval. The obtained plot is the reference against which we apply a standard linear interpolation method to estimate the values of the model parameters α,β,χ and δ that minimize the difference between the values of the two functions for each C_i.
In figure 4a the (not normalised) Precision and GRAmb are plotted for the test corpus. In figure 4b the normalised reference performance function and the "best fitting" scoring function are shown , with the estimated values of α,β,χ and δ.
While the reference function has a peak on the class set C_j with UB=55,000 and the score function assigns the maximum value to the class set C_k with UB=62,000, the performance of the sets in the range j-k is very similar. Table 1 shows the values of precision and global reduction of ambiguity in the range [j,k].

UB	Precision	GRAmb
55,000	0.752	0.338
56,000	0.758	0.347
59,000	0.748	0.360
60,000	0.734	0.378
61,000	0.731	0.386
62,000	0.733	0.368

Table 1 - Values of precision and global reduction of ambiguity for UB [55,000-62,000]

In evaluating the method, few aspects are worth un-

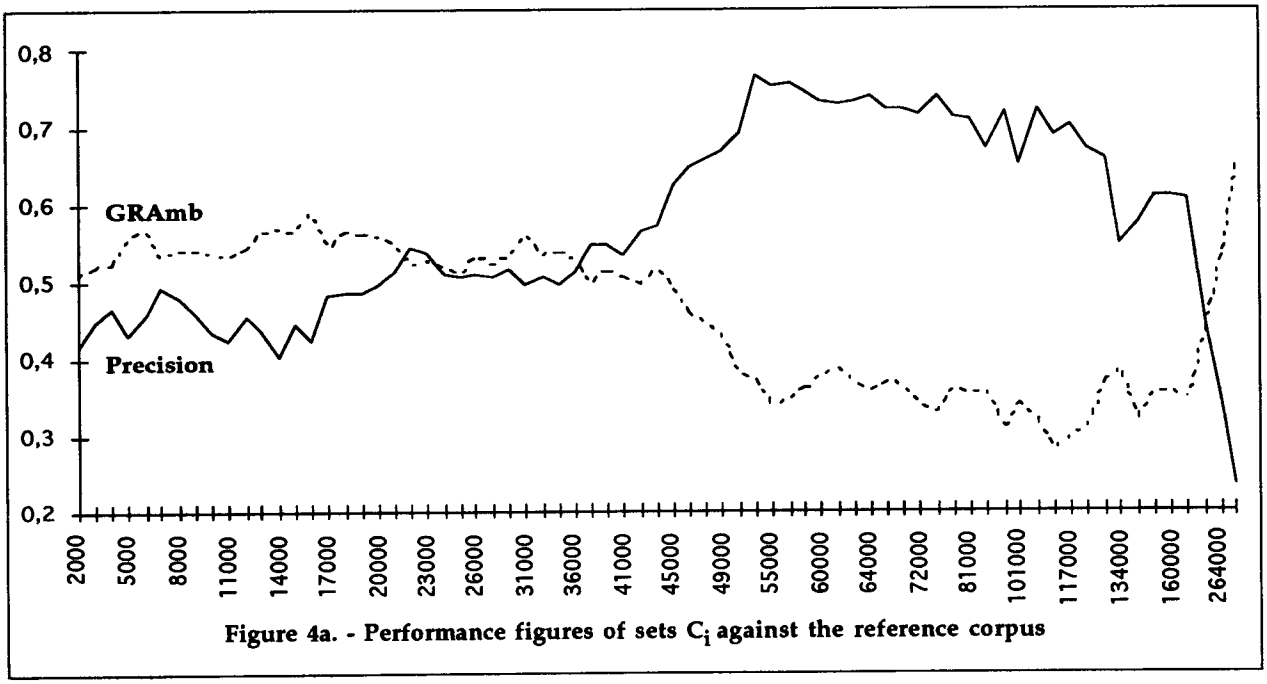

Figure 4a. - Performance figures of sets C_i against the reference corpus

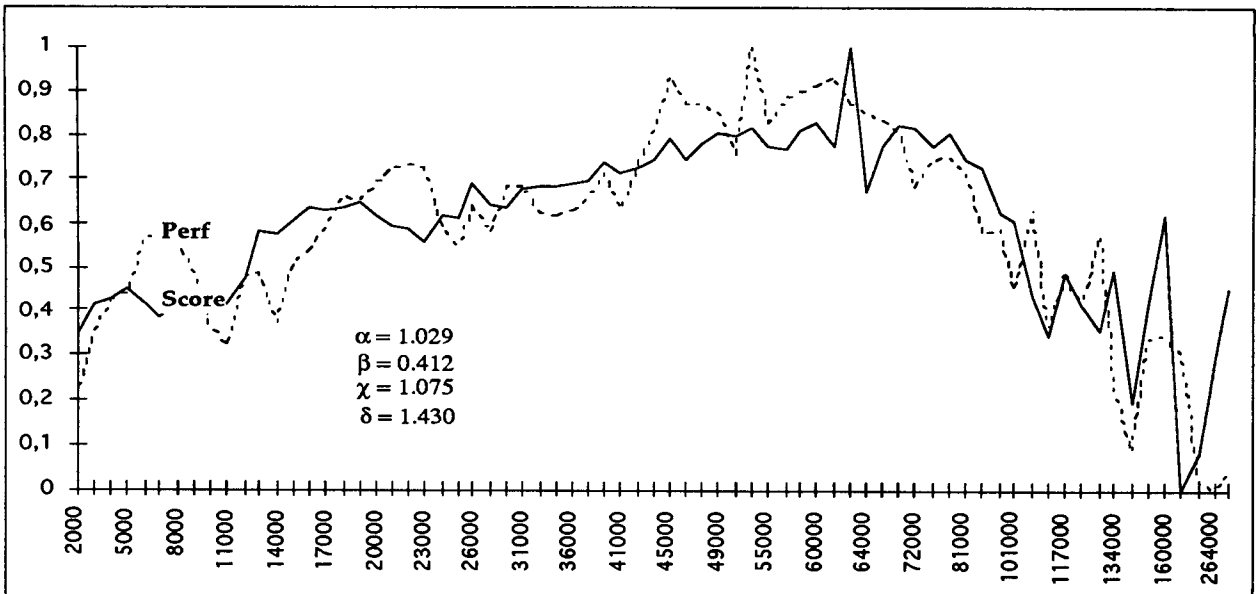

Figure 4b - Reference function and best-fitting Score function, with estimated parameters.

derlining

- The test corpus includes only 475 words of the over 10,000 in our learning corpus. This may well cause a shift of the reference scoring function, as compared with the "real" scoring function.
- In any case, figure 4a shows that the sets C_i have peak performances in the range 50,000-100,000. In this range, the precision is around 73-76%, and the reduction of ambiguity is around 35%, which are both valuable results. We also experimented that, by changing slightly the model parameters and/or the definitions of the four performance figures in the (1), in any case the peak performance of the obtained scoring function falls in the 50,000-100,000 interval, and the function stays high around the peak, with local *maxima*.
- In other domains (see a brief summary in the concluding remarks) for which we did not have a reference tagged corpus, we used (α=1 β=0,5 χ=1 δ=1) as model parameters in the (1), and still observed a scoring function similar in shape to that of figure 4b. Selected categories vary according to the domains, but the size of the best set stays around the 10-20 categories. Evaluation is of course more problematic due to the absence of a tagged reference corpus.

Therefore, we may conclude that *the method is "robust", in the sense that it correctly identifies a range of reasonable choices for the set of categories to be used, eventually leaving the final choice to a linguist.*

As for the WSJ corpus, a short analysis of the linguistic data may be useful. In figure 5 the 14 "best" selected categories for nouns are listed. Figure 6 shows four very frequent and very ambiguous words in the domain: *bank, business, market and stock*, with attached list of synsets as generated by WordNet, ordered from left to right by the increasing level of generality (leaf-sysnset leftmost). The senses marked with '*' are those that reach some of the categories (marked in bold in the figure) of the best-performing set, selected by the scoring function (1). For *bank* and *market*, we observed that the less plausible (for the domain) senses *ridge* and *market_grocery_store* are pruned out. The word *stock* retains only 5 out of 16 senses. Of these, the *gunstock* and *progenitor* senses should have been further dropped out, but there are 11 senses that are correctly pruned, like *liquid, caudex, plant*, etc. The word *business* still keeps its ambiguity, but the 9 subtle distinctions of WordNet are reduced to 7 senses.

```
 1.  person, individual, someone,
     mortal, human, soul
 2.  instrumentality, instrumentation
 3.  attribute
 4.  written_communication,
     written_language
 5.  message, content, subject_matter,
     substance
 6.  measure, quantity, amount, quantum
 7.  action
 8.  activity
 9.  group_action
10.  organization
11.  psychological_feature
12.  possession
13.  state
14.  location
```

Figure 5 - List of best-performing categories

Word: bank
Sense n.1: bank,side -> slope,incline,side -> ...
Sense n.2 (*): depository_financial_institution,bank,banking_concern,banking_company -> financial_institution,financial_organization -> institution,establishment -> **organization** -> ...
Sense n.3: bank -> ridge -> ...
Sense n.4: bank -> array -> ...
Sense n.5 (*): bank -> reserve,backlog,stockpile -> accumulation -> asset -> **possession**
Sense n.6 (*): bank -> funds,finances,monetary_resource,cash_in_hand,pecuniary_resource -> asset -> **possession**
Sense n.7: bank,cant,camber -> slope,incline,side -> ...
Sense n.8 (*): savings_bank,coin_bank,money_box,bank -> container -> **instrumentality,instrumentation** -> ...
Sense n.9: bank,bank_building -> depository,deposit,repository -> ...

Word: business
Sense n.1 (*): business,concern,business_concern,business_organization -> enterprise -> **organization** -> ...
Sense n.2 (*): commercial_enterprise,business_enterprise,business -> commerce,commercialism,mercantilism -> **group_action** -> ...
Sense n.3 (*): occupation,business,line_of_work,line -> **activity** -> ...
Sense n.4 (*): business -> concern,worry,headache,vexation -> negative_stimulus -> stimulation,stimulus,stimulant,input -> information -> cognition,knowledge -> **psychological_feature**
Sense n.5 (*): business -> aim,object,objective,target -> goal,end -> content,cognitive_content,mental_object -> cognition,knowledge -> **psychological_feature**
Sense n.6: business,business_sector -> sector -> ...
Sense n.7 (*): business -> business_activity,commercial_activity -> **activity** -> ...
Sense n.8: clientele,patronage,business -> people -> ...
Sense n.9 (*): business,stage_business,byplay -> acting,playing,playacting,performing -> **activity** -> ...

Word: market
Sense n.1: market -> class,social_class,socio-economic_class -> ...
Sense n.2: grocery_store,grocery,market -> marketplace,mart -> ...
Sense n.3 (*): market,marketplace -> **activity** -> ...
Sense n.4 (*): market,securities_industry -> industry -> commercial_enterprise -> enterprise -> **organization** -> ...

Word: stock
Sense n.1 (*): stock -> capital,working_capital -> asset -> **possession**
Sense n.2 (*): stock,gunstock -> support -> device -> **instrumentality,instrumentation** -> ...
Sense n.3: stock,inventory -> merchandise,wares,product -> ...
Sense n.4 (*): stock_certificate,stock -> security,certificate -> legal_document,legal_instrument,official_document,instrument -> document,written_document,papers -> writing,written_material -> **written_communication,written_language** -> ...
Sense n.5 (*): store,stock,fund -> accumulation -> asset -> **possession**
Sense n.6 (*): stock -> progenitor,primogenitor -> ancestor,ascendant,ascendent,antecedent -> relative,relation -> **person,individual,someone,mortal,human,soul** -> ...
Sense n.7: broth,stock -> soup -> ...
Sense n.8: stock,caudex -> stalk,stem -> ...
Sense n.9: stock -> plant_part -> ...
Sense n.10: stock,gillyflower -> flower -> ...
Sense n.11: Malcolm_stock,stock -> flower -> ...
Sense n.12: lineage,line,line_of_descent,descent,bloodline,blood_line,blood,pedigree,ancestry,origin, parentage,stock -> genealogy,family_tree -> ...
Sense n.13: breed,strain,stock,variety -> animal_group -> ...
Sense n.14: stock -> lumber,timber -> ...
Sense n.15: stock -> handle,grip,hold -> ...
Sense n.16: neckcloth,stock -> cravat -> ...

Figure 6. Selected synsets for the words *bank, business, market* and *stock.*

5 Concluding Remarks

It has already been demonstrated in (Basili et al, 1996) that tagging a corpus with semantic categories triggers a more effective lexical learning. However, overambiguity of on-line thesaura is known as the major obstacle to automatic semantic tagging of corpora. The method presented in this paper allows an efficient and simple selection of a flat set of *domain-tuned* categories, that dramatically reduce the initial overambiguity of the thesaurus. We measured a 73% precision in reducing the initial ambiguity, and a 37% global reduction of ambiguity. Significantly, our method selects a limited number of categories (10-20, depending upon the learning corpus and the model parameters), out of the initial 47,110 leaf-synsets of WordNet[4].

We remark that our experiment is *on large*, meaning that we automatically evaluated the performance of the model on a large set of nouns taken from the Wall Street Journal. Most sense disambiguation or semantic tagging methods evaluate their performances manually, against few very ambiguous cases, with clear distinctions among senses. Instead, WordNet draws very subtle and fine-grained distinctions among words. We believe that our results are very encouraging.

The model parameters for category selection has been tuned on SEMCOR, but a correctly tagged corpus is not strictly necessary. In our experiments, we applied a scoring function similar to that obtained for the Wall Street Journal to two other domains, a corpus of Airline reservations and the Unix handbook. We do not discuss the data here for the sake of space. The method constantly selects a set of categories at the medium-high level of generality, different for each domain. The selection "seems" good according to our linguistic intuition of the domains, but the absence of a correctly tagged corpus does not allow a large-scale evaluation.

In the future, we plan to demonstrate that the method proposed in this paper, besides reducing the overambiguity of on-line thesaura, improves the performance of lexical learning methods that are based on semantic tagging, such as PP disambiguation, case frame acquisition and and sense selection, with respect to a non-optimal choice of semantic categories.

6 Acknowledgements

The method presented in this paper has been developed within the context of the ECRAN project LE 2110, funded by the European Community. One of the main research objectives of ECRAN is *lexical tuning*, being *semantic tagging* and *sense disambiguation* two important and preliminary objectives. This paper approached the problem of domain-appropriate semantic tagging.

We thank Christian Pavoni who developed much of the software used in this experiment, as well as all our partners in the ECRAN project.

[4]We used WordNet in this experiment, because it is now very popular among scholars in lexical statistics, but clearly our method could be applied to any on-line taxonomy or lattice.

References

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

(Basili et al, 1992) Basili, R., Pazienza, M.T., Velardi, P., "Computational Lexicons: the Neat Examples and the Odd Exemplars", *Proc. of Third Int. Conf. on Applied Natural Language Processing*, Trento, Italy, 1-3 April, 1992.

(Basili et al, 1996) Basili, R., M.T. Pazienza, P. Velardi, An Empyrical Symbolic Approach to Natural Language Processing, *Artificial Intelligence*, August 1996

(Basili et al, 1996b) R. Basili, R., M.T. Pazienza, P. Velardi, Integrating general purpose and corpus-based verb classification, *Computational Linguistics*, 1996

(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*

(Dagan et al, 1994)Dagan I., Pereira F., Lee L., Similarly-based Estimation of Word Co-occurrences Probabilities,*Proc. of ACL*, Las Cruces, New Mexico, USA, 1994.

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

(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.

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

(Grishman and Sterling, 1994) R. Grishman, J. Sterling, Generalizing Automatically Generated Selectional Patterns, *Proc. of COLING '94*, Kyoto, August 1994.

(Pereira et al., 1993) Pereira F., N. Tishby, L. Lee, Distributional Clustering of English Verbs, *Proc. of ACL*, Columbus, Ohio, USA, 1993.

(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

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

(Utsuro et al, 1993) T. Utsuro, Y. Matsumoto and M. Nagao, "Verbal case frame acquisition from Bilingual Corpora" *Proc. of IJCAI*, 1993