## Controlled Bootstrapping of Lexico–semantic Classes as a Bridge between Paradigmatic and Syntagmatic Knowledge: Methodology and Evaluation

### Paolo Allegrini, Simonetta Montemagni, Vito Pirrelli

Istituto di Linguistica Computazionale – CNR Via Alfieri 1 - Pisa 56010 - ITALY {allegrip, simo, vito}@ilc.pi.cnr.it

#### Abstract

Semantic classification of words is a highly context sensitive and somewhat moving target, hard to deal with and even harder to evaluate on an objective basis. In this paper we suggest a step–wise methodology for automatic acquisition of lexico–semantic classes and delve into the non trivial issue of how results should be evaluated against a top–down reference standard.

### 1. Introduction

Assessing the semantic similarity between words plays an instrumental role in a variety of Natural Language Processing (NLP) tasks, including syntactic and word sense disambiguation, selection of appropriate translation equivalent, assessment of lexical cohesion in texts for automatic summarisation, query expansion and document indexing in Information Retrieval. Semantic similarity, however, is a pre– theoretical notion, and its elusive nature escaped many attempts to make it algorithmic and operationally useful.

Two approaches to the problem have been prevalent. Semantic similarity is "paradigmatic", i.e. it is based on hand–crafted taxonomic repositories of lexico–semantic information such as WordNet. Semantic similarity is "syntag-matic", that is it is grounded on the assumption that words entering into the same contextual relation with other words are semantically similar (see, among others, Pereira and Tishby 1992; Pereira *et al.* 1993; Rooth 1995; Rooth *et al.* 1999). Paradigmatic approaches are, in general, insensitive to the basic fact that similarity changes as context and topic change. On the other hand, syntagmatic approaches fail to make it explicit what sort of contextual sameness has a bearing on semantic similarity.

The two approaches are clearly complementary, but it is still unclear whether they could be integrated in any meaningful way. Using semantic similarity to cluster words into semantic classes strikes us as a way to evaluate this possibility objectively. Since available taxonomies provide a top-down classification of words, there is wide room for assessing to what extent a top-down typology meets a bottom-up classification based on distributional similarity. With this purpose in mind, we propose here i) an algorithm for controlled, explorative bootstrapping of lexicosemantic classes from corpus and dictionary data and ii) an evaluation protocol for projecting acquired classes against hand-crafted typologies. The algorithm is explorative since clusters are built incrementally from data types. As evaluation is incremental too, we are in a position to observe what goes wrong with the clustering, and to point out reasons for that: e.g. domain-specific world knowledge, typology of linguistic knowledge etc. We hope this could shed light on the interaction between paradigmatic and syntagmatic lexico-semantic approaches to the problem.

The first part of the paper (sections 2., 3. and 4.) de-

scribes methodology and techniques developed for the acquisition of lexico–semantic classes. The second part (section 5.) focuses on evaluation of acquired data.

### 2. CLASS: general principles

Identification of semantic clusters of words is carried out by CLASS (CLustering through Analogy–based Semantic Similarity) on the basis of analogy–based semantic similarity measures.

The main relevant assumptions underlying the proposed method can be summarized as follows: i) classification is highly **context sensitive**; ii) still classification is **asymmetric** (unlike the notion of semantic similarity which is in fact **symmetric**): this means that classifying a lexical unit in context is not the same as classifying its context; iii) classification is **multidimensional**: *i.e.* changing semantic dimension has an obvious impact on classification: still the empirical hypothesis has to be tested that dimensions may have a finite order of magnitude; for this purpose we need truly explorative, bootstrapping techniques; iv) as classification is a moving target, we have to make provision for different protocols to evaluate results; objective assessment of obtained classification is thus a crucial issue.

CLASS does not assume any preexisting semantic classification, but grounds semantic generalizations on controlled distributional evidence. In particular, semantic classes are derived from the distributional analogies observed in a Knowledge Base (KB) of functionally annotated word co–occurrence patterns. Intuitively, two words are taken to be distributionally similar if they tend to be in complementary distribution relative to the same syntactic contexts (or co–occurrence patterns), where criteria for sameness involve identity of both (i) the words making up the context (and obviously of the associated part of speech) and (ii) the syntactic functions holding between them. We use the notion of "analogical proportion" to give this intuition a more formal attire.

#### 2.1. Analogical Proportions

For the present purposes, a word co-occurrence pattern is defined as a pair of functionally-annotated words, e.g. a noun, a verb and the syntactic function holding between the two.<sup>1</sup> This can be represented formally as the triple  $[v_k; n_i; f_m]$ . An analogical proportion is then a quadruple of functionally annotated pairs resulting from the combination of any two nouns  $n_i$  and  $n_j$  with any two verbs  $v_k$  and  $v_t$  such as (1) holds:

where terms along the two diagonals can swap place in the proportion, and identity of subscript indicates identity of values.

Some properties of (1) are worth emphasizing here. First, it does not require that the same syntactic function hold between all pairs, but only that functions be pairwise identical: the syntactic functions associated with different verbs in the same analogical proportion are not to necessarily be identical. On the other hand, identity at the level of syntactic function is conditional on identity of verbs, i.e. identical verbs should be associated with the same syntactic function. This means that, for any given verb  $v_k$ , we are treating  $[v_k; n_i; f_m]$  and  $[v_k; n_i; f_n]$  as two different syntactic contexts if  $f_m \neq f_n$ . This responds well to our intuition about context identity. We call this constraint the "same– verb–same–function" principle.

In the remainder of the paper, we will take (1) as the operational definition of the analogical proportion to be used for semantic clustering, in particular for inferring the semantic similarity of distributionally equivalent words. This definition requires that two words be used interchangeably in at least **two** different contexts.

### 2.2. Weighting analogical proportions

Not all APs are equally relevant to an assessment of the semantic similarity between words. Some APs are certainly less significant than others. For example, that shower and care are both found as objects of take is less solid evidence of their semantic similarity than the fact that - say - both *pipe* and *cigarette* are objects of *smoke*. In order to reflect this graded level of relevance, each AP can be assigned a score. Intuitively, the score should be sensitive to the fact that a word which keeps company with a small set of collocates - i.e. which is lexically choosy - is semantically heavier than words co-occurring with a high number of collocates, as is the case of so-called "light verbs" like *take*. The most highly valued APs are then those containing word co-occurrence pairs involving semantically heavier words. Hence, heuristically, the desired score should define a function whose values are low when semantically light words form part of a given AP, and high when semantically heavier words are involved. We will hereafter refer to this score as a Correlation Score.

We define the correlation score of an AP as a monotonically decreasing function of the **type frequency** of the verbs and nouns forming AP. Since the score should reflect the extent to which two words are similar, and since similarity is a symmetric relation, it is reasonable to assume that the score be symmetric too. We also make the assumption that the way words contribute to the calculation of the score is independent of their category: i.e., the semantic weight of nouns and verbs affects the score of an analogical proportion in a symmetric way.<sup>2</sup>

The current version of CLASS works on a KB of pair types only, with no information about the token–frequency of each word pair in a corpus. This means that each pair is associated with the same amount of information, as it occurs only once. Unlike pairs, words can participate in more than one pair of KB, so that their conveyed information varies proportionally to the number of occurrences of each word a in our KB, corresponding to the word type frequency  $\omega(a)$ . More concretely, assuming that the word a is a verb, the information content I associated with a is:

$$I_a = -\sum_{[a;n]\in KB} \log_2 p([a;n]) = \sum_{[a;n]\in KB} constant \propto \omega(a),$$
(2)

where p([a; n]) is the probability of finding the pair [a; n] in KB, proportional to the (constant) frequency of the pair. An analogue of eq. 2, with [v; a] replacing [a; n], holds when a is a noun. Intuitively, eq. 2 says that the sum of some constant terms (whose value is independent of the word a) is proportional to the number of the terms in the sum, and thus, ultimately, to the word type frequency  $\omega(a)$ .

It is then natural to define the correlation score of AP,  $\sigma(v_1, v_2, n_1, n_2)$ , as follows:

$$\sigma(v_1, v_2, n_1, n_2) = \frac{\omega(AP)}{\omega(v_1)\omega(v_2)\omega(n_1)\omega(n_2)}, \quad (3)$$

where  $\omega(AP)$  indicates the type frequency of the proportion: namely 1 if the proportion is attested in KB (all 4 pairs are present) or 0 if it is not.<sup>3</sup>

To sum up, equation (3) is compatible with our assumptions of both order symmetry (invariance to the order in which two words are considered) and categorial symmetry (nouns carry the same weight as verbs). Furthermore, it embodies the heuristics according to which semantically light words contribute less to the correlation score than semantically heavier words do. Finally, it can be given an informational (entropic) interpretation. The adopted measure is the most natural candidate score for weighting analogical proportions if only type frequencies can be used (either because token frequencies are not available or because they cannot be trusted due to data sparseness).

### **3.** Word Clustering by CLASS

Word clustering goes through different steps:

<sup>&</sup>lt;sup>1</sup>In what follows, we will keep talking about functionally annotated verb–noun patterns; it should be appreciated, however, that analogical proportions can in principle be established among word patterns of any type.

<sup>&</sup>lt;sup>2</sup>This is not to mean that **classification** of both nouns and verbs in the same context is symmetric. As argued in (Allegrini *et al.* 2000) maximization of the similarity of nouns (verbs) may cause minimization of the similarity of their accompanying verbs (nouns). A class where a maximum of noun similarity correlates with a maximum of verb similarity can be uninformative, as is the case of poorly selective verbs such as *give, find, get* etc.

 $<sup>^{3}</sup>$ A formal derivation of eq. (3) is given in (Allegrini *et al.* 2000).

- 1. search space carving;
- 2. identification, validation and ranking of potential semantic clusters;
- 3. cluster lumping.

In what follows, each step is described in detail.

### 3.1. Step 1: Search Space Carving

Given a pool P of words to be clustered semantically, the first step of CLASS consists in carving out a relevant portion of KB. To give a practical example, let P consist of the following set of nouns:

appesantimento 'increase in weight', crescita 'growth', flessione 'decrease', guaio 'trouble', problema 'problem', rialzo 'rise', ridimensionamento 'reduction', ritardo 'delay', turbolenza 'turbulence'.

The set represents all object collocates of the Italian verb *causare* 'cause' (also referred to as the *headword*), as they are found in the underlying knowledge base.

In our KB guaio 'trouble' is used as (i) subject of the verb *capitare* 'happen', (ii) prepositional complement of the verb *incappare* 'run into', (iii) object of *causare* 'cause'. The resulting carved search space will then include, besides all verb–object pairs involving *causare*, all pairs involving the verbs co-occurring with nouns in P with the same syntactic function as the one attested with nouns in P (see the same–verb–same–function constraint), e.g. all verb–subject pairs relative to *capitare* and all verb–complement (namely prepositional complement headed by the preposition *in*) pairs of *incappare*.

A carved search space can thus include word cooccurrence pairs involving different types of dependency. This is felt useful since it makes it possible to mutually relate contextual similarities based on different syntactic functions, and eventually extrapolate a similarity grounded on one type of dependency relation to another "syntactically asymmetric" context (see 2.1. above).<sup>4</sup>

# **3.2.** Step 2: Identification and ranking of potential semantic clusters

Identification of potential semantic clusters operates within the search space carved out at the previous stage, and requires preliminary identification of so-called "substitutability islands" (SIs), that is *n*-tuples of verbs and nouns, where all nouns are attested as co-occurring with all verbs with the same syntactic function. For instance, the verbs *causare* 'cause' and *incappare* 'run into' form a substitutability island together with the nouns *guaio* 'trouble' and *problema* 'problem', as the four words combine together in word co-occurrence patterns where the "sameverb-same-function" constraint holds (eq. (1) above).<sup>5</sup> The APs described in section 2.1. can accordingly be seen

· · · · · · · · · · · · · · · · · · ·		
FLESSIONE, RIALZO }	:	CAUSARE / 0, REGISTRARE / 0
{CRESCITA, FLESSIONE}	4	CAUSARE / 0, EVIDENZIARE / S,
		MEDIARE / 0, MOSTRARE / 0,
		PRESENTARE / S, PRESENTIRE / O
		REGISTRARE / 0, REGISTRARE / S
{CRESCITA, GUAID}	4	CAUSARE / O, PROVOCARE / O
{CRESCITA, PROBLEMA}	:	AVERE / S, CAUSARE / O, EVIDENZIARE / O
		PORRE / S, PRESENTARE / S
{CRESCITA, RITARDO}	:	CAUSARE / 0, USARE / S
{FLESSIONE, PROBLEMA}	:	CAUSARE / 0, PRESENTARE / S, STARE / S
{FLESSIONE, RIALZO}	:	CAUSARE / 0, REGISTRARE / 0
		REGISTRARE / S, SUBIRE / O
{GUAIO, PROBLEMA}	:	CAUSARE / 0, CAVARE - SI / S, INCAPPARE / S
{RIDIMENSIONAMENTO, RITARDO}	1	CAUSARE / D, GIUSTIFICARE / D

√APPESANTIMENTO CRESCITA

Figure 1: Some SIs relative to the nouns in P.

as minimal substitutability islands, where the *n*-tuple consists of one pair of verbs and one pair of nouns only. Bigger substitutability islands presuppose as many APs as different attested quadruples of contextually interchangeable nouns and verbs. Nouns belonging to the same SI form a potential semantic cluster. In Fig.1, we give a sample of semantic classes of nouns (between curly brackets) based on the notion of substitutability island. Beside each noun class (in curly brackets) we give, after a colon, a list of the verbs forming a SI with the nouns. The suffix following each verb indicates the type of dependency holding between nouns in the set and each verb ('S' for subject, 'O' for object).

Note that extracted classes include both semantically coherent clusters as well as heterogeneous ones. Consider, for instance, the cluster {*crescita problema*} 'growth problem' in Fig.1, whose nouns do not appear to have much in common from a semantic point of view besides their being both objects of the verb *causare*. Intuitively more semantically plausible clusters for *crescita* are {*crescita flessione*} 'growth, decrease' and {*appesantimento crescita flessione rialzo*} 'increase in weight, growth, decrease, rise'. By the same token, *problema* would form a semantically more coherent pair with a word like *guaio* 'trouble'.

The output in Fig.1 shows an excessive degree of fragmentation of the semantic space, the average size of a potential semantic class being only slightly larger than 2. This is counterbalanced by the existence of significant intersections among classes and suggests that there is room for improvement: namely, bigger classes can be formed by merging existing intersecting clusters. We will refer to this step as "class lumping" (see section 3.3. below). Before going into that, however, a validation stage is in order to rank the extracted semantic clusters according to their significance; this is done by assigning each potential semantic class a relevance score.

#### 3.2.1. Class score

Our measure for scoring potential semantic classes naturally stems from the correlation score  $\sigma$  described in section 2.2. Let us start with a class formed by a pair of nouns. It is straightforward to define the correlation score  $\sigma(n_1, n_2)$  of

<sup>&</sup>lt;sup>4</sup>On the effectiveness and usefulness of relating syntactically asymmetric contexts to infer semantic similarity, the interested reader is referred to (Federici *et al.* 1997).

<sup>&</sup>lt;sup>5</sup>In particular, *guaio* and *problema* are both objects of *causare* and prepositional complements (headed by the preposition *in* 'in') of *incappare*.

7.04509e-05{GUAIO,PROBLEMA}
7.01459e-05{RIDIMENSIONAMENTO,RITARDO}
4.65858e-05{CRESCITA,FLESSIONE}
1.75699e-05{FLESSIONE,RIALZO}
9.49509e-06{APPESANTIMENTO,CRESCITA,FLESSIONE,RIALZO}
1.88964e-06{CRESCITA,GUAIO}
1.19814e-06{CRESCITA,RITARDO}
8.84254e-07{CRESCITA,PROBLEMA}
6.7141e-07{PLESSIONE,PROBLEMA}

Figure 2: Nine top-most scored noun clusters

a pair of nouns as

$$\sigma(n_1, n_2) \equiv \frac{1}{2} \sum_{v_1, v_2 \neq v_1} \sigma(v_1, v_2, n_1, n_2), \qquad (4)$$

where the verbs  $v_1$  and  $v_2$  are both attested in KB. The summation over the different verb indexes in (4) counts every couple twice, which is balanced by factor 1/2. Since we have only 2 nouns the summation can also be interpreted as a sum *over all possible analogical proportions*. The definition of weight of a class in the CLASS experiments reported below is actually

$$\sigma(C) \equiv A(C) \cdot \sum_{\forall AP \in KB} \sigma(v_1, v_2, n_1, n_2), \qquad (5)$$

where  $C = \{n_1, \cdots, n_k\}$  is a class of k nouns,

$$A(C) = \frac{1}{k}.$$
 (6)

In our experiment the influence of k, which is responsible for breaking the noun/verb symmetry in our treatment, was tested by forcing it to unity and repeating the experiment. Since the difference in the results is not dramatic (in fact almost negligible), the results reported in the following sections only refer to the definition of A(C) in eq. 6.

Potential semantic classes are then ranked by decreasing values of  $\sigma(C)$  in Fig. 2, where  $\sigma(C)$  is calculated according to eq. 5. It should be appreciated that the obtained ranking in the list considerably meets our intuition: groupings such as {guaio problema} 'trouble, problem', {crescita flessione} 'growth, decrease', {appesantimento crescita flessione rialzo} 'increase in weight, growth, decrease, rise' are assigned a higher score with respect to the classes {crescita problema} 'growth, problem' and {flessione problema} 'decrease, problem' which appear to be ranked at the bottom in the list.

#### 3.3. Step 3: Centroid identification and lumping

This step operates on ranked lists of potential semantic clusters to form final semantic classes. This is carried out in two different stages.

The first stage is identification, among the extracted potential semantic clusters, of class "centroids", that is sets of tightly associated words representing the most typical semantic cores around which final semantic classes are eventually constructed. We make the basic assumption that centroids are disjunctively defined: that is, there exists no pair 7.04509e-05 {GUAIO PROBLEMA} 7.01459e-05 {RIDIMENSIONAMENTO RITARDO} 4.65858e-05 {CRESCITA FLESSIONE}

#### Figure 3: Selected centroids

{ (GUAIO PROBLEMA) | TURBOLENZA } { (RIDIMENSIONAMENTO RITARDO) } {
(CRESCITA FLESSIONE) | RIALZO | APPESANTIMENTO }

#### Figure 4: Lumped classes

of intersecting centroids. The best possible selection of such cores will then include non–intersecting clusters with the highest possible cumulative score. In practice, the first centroid corresponds to the class with the topmost score. The second best centroid is the highest class with no intersection with the first centroid, and so on (the *i*-th centroid is the highest class with no intersection with the first centroid second with the first i-1 centroids) until all clusters in the rank are used up. This squares with the assumption that centroids represent the principal meaning components of the verb headword and that these components are mutually orthogonal. Centroid selection is evaluated separately, as shown in section 5.

The second stage consists in lumping classes together around the identified centroids. A variety of different strategies can in principle be adopted for lumping. Here below we will show results obtained thus: i) for each centroid  $c_k$ , we first select classes intersecting with  $c_k$  only and lump them with  $c_k$ ; ii) at the end of step i), each outstanding noun N is assigned to the centroid with the largest intersection with the classes containing N. However simple, this choice has both practical and theoretical advantages. First, it avoids building intersecting classes, and, in turn, dispenses with ambiguities in evaluating the output of automatic classification against a manually crafted reference classification with no intersecting semantic classes (see section 5.). Secondly, it yields a set of maximally orthogonal and semantically coherent noun classes, under the assumption that these classes highly correlate with the principal meaning components of the verb head of which input nouns are objects.

Let us exemplify the two different stages of step 3 starting from the ranked list of potential semantic classes reported in Fig. 2 above. The potential semantic classes selected as centroids are reported in Fig. 3. Extended semantic classes – or "lumped classes" in our terminology – are formed through set union of each selected centroid and the clusters intersecting it. More specifically, the lumping strategy exemplified in Fig. 4 follows a) above. In Fig. 4, the final class {*guaio problema turbolenza*} is formed by merging the centroid {*guaio problema*} 'trouble, problem' with the intersecting class {*problema* turbolenza} 'problem, turbulence' through set union. <sup>6</sup> Added elements which come from different intersecting classes are separated by " ]", as illustrated in the third line of Fig. 4.

Incidentally, it should be noted that the lumped classes of Fig. 4 appear to capture fine grained semantic distinctions. For instance, a distinction is made between **incre**-

<sup>&</sup>lt;sup>6</sup>This class does not appear in Fig. 2 due to its low score.

**mental events** or results of incremental events, which presuppose a scalar dimension: this is the case of {*crescita flessione rialzo appesantimento*} 'growth, decrease, rise, increase in weight'; and **rescheduling events**, where a change occurs with respect to a previously planned event or object: see the class {*ridimensionamento ritardo*} 'reduction delay'.

Distinctions like these ones are often eluded in the taxonomical organization of a lexicon.

### 4. CLASS at work

Clustering experiments have been carried out through different settings, whose main parameters of variation are: i) type of input words to be semantically classified, and ii) configuration of the KB against which input data are projected. In what follows, these two points will be discussed in detail in sections 4.1. and 4.2. respectively.

#### 4.1. Input to CLASS

So far, we focused on noun clustering, with particular emphasis on object collocates of a given verb. This was due to the type of task we set ourselves to, namely semantic annotation of complement positions in verb subcategorisation frames. Nonetheless, nothing opposes in principle to clustering words other than nouns, given a KB of word cooccurrence patterns containing the words in question.

Although noun sets can be carved out according to completely different criteria (e.g. they can be defined as the leaves of a specific branch in a noun taxonomy), we believe that starting from the collocates of a verbal head yields local lexico–semantic classifications which are those urgently needed in NLP tasks based on the assessment of semantic similarity: in fact, semantic similarity is mainly a context– sensitive notion (Miller and Charles, 1991). Among the practical advantages of local classifications we should at least mention the following two: choice of a verb head as a perspectivizing factor considerably reduces the possibility that the same polysemous object collocate is used in different senses with the same verb; furthermore, the resulting clusters can give information about the senses, or meaning facets, of the verb head.

In all experiments described here, object collocates were extracted from the "Italian SPARKLE Reference Corpus", a corpus of Italian financial newspapers of about one million word tokens used as data source in the phase of automatic frame acquisition (see (Carroll *et al.* 1997; Federici *et al.* 1998)). In that phase, we acquired the sets of object collocates of 20 Italian verbs, chosen among the most frequent translational equivalents of the 20 English verbs forming the test–bed of SPARKLE evaluation protocol.<sup>7</sup> The object collocates of each test verb were automatically extracted from the entire SPARKLE reference corpus during the subcategorisation induction stage: for each position

of an acquired frame, the lexical acquisition system keeps also track of the lexical fillers instantiating it.

Additional experiments were also carried out for the same verbs on the basis of the typical object collocates attested both in definitions and example sentences of two dictionaries, a bilingual Italian–English dictionary (Collins 1985) and a monolingual one (Garzanti 1984).

### 4.2. KB configuration

Acquisition experiments with CLASS were carried out on different KB configurations, varying with respect to a number of different parameters:

- 1. internal composition;
- 2. data source;
- 3. size.

Concerning point 1, KBs with different types of word cooccorrence patterns were used, namely:

- a) verb-object patterns (e.g. *causare-problema*/O 'cause-problem');
- b) verb-subject and verb-object patterns (including also patterns such as *capitare-problema*/S 'occurproblem');
- c) verb-subject, verb-object, and verbprepositional complement patterns (also including patterns such as *incappare-problema*/IN 'runproblem/into').

Note, incidentally, that KBs with the typology of patterns decribed in b) and c) make it possible to resort to APswith asymmetric syntactic constructions. In other words, it is possible to infer the contextual similarity relative to a certain syntactic function on the basis of the attested contextual similarity relative to another function.

The second parameter (point 2) refers to the type of linguistic resources used to build the Knowledge Base underlying CLASS. Basically, two different knowledge sources have been experimented with: i) dictionaries, both bilingual and monolingual ones; ii) unrestricted texts, namely the SPARKLE financial corpus used for acquisition. The two sources reflect two different types of linguistic usage: typical examples of use of a given word in the case of dictionaries, and actual usage of words in the case of running corpora. In principle, different knowledge sources should exemplify a different typology of senses. Dictionaries usually testify all possible senses of a given word. Therefore, typical word collocates acquired from dictionaries tend to cover the entire range of possible senses of a headword. On the other hand, unrestricted texts reflect actual usage and possibly bear witness to senses which are relevant to a specific domain only. For more information on the different results obtained by using different KBs the reader is referred to section 5. below. Experiments were also carried out by combining data from both dictionaries and corpora. Comparative results for these different settings are illustrated in section 5.

As to point 3 above, inclusion of different configurations of patterns give rise to KBs of different size, ranging from about 18,000 different verb–noun pairs (verb–object

<sup>&</sup>lt;sup>7</sup>The test verbs are: AGGIUNGERE 'add', AIUTARE 'help', ASPETTARE 'expect', CAMBIARE 'change', CAUSARE 'cause', CHIEDERE 'ask', CONSIDERARE 'consider', DARE 'give', DE-CIDERE 'decide', FORNIRE 'provide', MUOVERE 'move', PER-METTERE 'allow', PORTARE 'bring', PRODURRE 'produce', SCEGLIERE 'choose', SENTIRE 'feel', STABILIRE 'establish', TAGLIARE 'cut', TERMINARE 'end', TROVARE 'find'.

patterns only) to 43,000 pairs, with a KB containing verb– subject, verb–object and verb–prepositional complement patterns.

### 5. Evaluation of acquired results

Correctness of the semantic classes acquired by CLASS was tested through different evaluation procedures, both intrinsic and extrinsic. An intrinsic evaluation protocol, carried out against different types of pre–existing semantic classifications, is reported below. As to extrinsic evaluation, induced semantic classes were used to discriminate attested verb-object pairs from other made-up pairs containing the same verb (Pereira *et al.* 1993): description of results obtained through this evaluation protocol will be reported elsewhere. Finally, inferred semantic classes were also tested in the framework of other NLP tasks: namely, subject/object disambiguation in Italian (Montemagni *et al.* 1996) and word sense disambiguation (Federici *et al.* 1997; Federici *et al.* 1999a); Federici *et al.* 1999b).

### 5.1. Gold standard

For the present purposes, a gold standard is a partition of the set P of input words into semantic classes. A set partition requires that each word  $w_i$  in P may belong to one and only one identified partition class. There exists no empty subclass, but a subclass can consist of one element only. No word  $w_i$  in P can be discarded or left unclassified.

The choice of modeling a semantic classification in terms of set partition is justified in the light of the internal composition of P which is carved out as the set of collocates of a given verb: this reduces the problem of lexical ambiguity of members of P considerably, thus making plausible the assumption that the resulting semantic classes are defined disjunctively. Moreover, careful analysis of the output results of CLASS shows that contextual selection of P reduces the problem of multidimensional lexical similarity to a considerable extent. The problem has to do with the fact that the meaning of words is inherently multidimensional (not to be confused with polysemy or ambiguity), so that word senses which lie close along one dimension can turn out to be placed at opposite ends along another (orthogonal) dimension. Montemagni and Pirrelli (1998) argue that this phenomenon has significant repercussions on NLP applications and that it represents a stumbling block in the design of word taxonomies (which are typically monodimensional). They also show that this difficulty can be reduced, if semantic classifications are local or contextually salient, that is if they are defined in relation to a pool of contextually-determined senses, rather than in terms of a context-free selection of word senses taken in isolation. In practice this means that each word in P is assumed, for the purposes of evaluation, to have one contextuallydetermined sense only.

Validation at this stage is also intended to test how successfully a semantic classification of collocates can mirror the lexical preferences of a verb head on its arguments, and, eventually, the extent to which a difference in lexical preferences is indicative of a difference in word meaning. If confirmed the results would show that a semantic classification of verb collocates can tell us a lot about the degree

of ambiguity/polysemy of the head verb in question.

A principled difficulty with this approach to validation is represented by the classical egg–and–chicken problem: how is it possible to create an *a priori* suitable gold standard to evaluate a classification which is by definition data– driven? We do not have a final answer to this question. We just tried to minimize the degree of inevitable subjectivity involved in this process, by defining a semantically coherent class of nouns as the set of collocates of one sense of a verbal headword, as attested in a reference monolingual dictionary (Garzanti 1984).

The gold standard was thus built as follows: given a verb  $v_k$  and a set P of its noun collocates, a human annotator was asked to decide, for each noun in P, which of the possible senses of  $v_k$  attested in the reference dictionary may have selected it. Nouns which are selected by the same verb sense are assigned the same semantic subclass. This is not always easy: i) in some cases, the same noun could have been associated with different verb senses, since the difference between the involved senses did not entail a significant difference in selectional preferences; ii) in some other cases, an extra sense should have been added to the list of senses attested in the dictionary, when the particular context expressed by a verb-noun pair was apparently not covered by any of the attested senses. In practice, for this run of tests we decided to stick to the dictionary predefined senses. Hence, case i) was handled by looking for the most typical verb sense selecting for the noun to be classified; in case ii), the most plausible - but possibly not fully appropriate - sense was identified among the attested ones. The reference classification built along these lines will be referred to as gold standard. Noun classes in the gold standard are still fairly coarse-grained. Their semantic cohesion is nonetheless guaranteed by their being collocates of the same word sense.

### 5.2. Evaluation parameters

Careful analysis of preliminary results taught us that assessment of set partitions against a manually–crafted gold standard requires a multivariate analysis of output data, since there is more than one significant parameter for evaluating how well the system is meeting our requirements. In what follows, we overview these parameters and give reasons for their use.

### 5.2.1. Centroid evaluation

The first evaluation parameter is concerned with identification of centroids. Given the classes in the "gold standard" and the inferred set of centroids, recall is defined as the ratio between the number of gold classes which completely include at least one inferred centroid and the number of gold classes. Precision is the ratio between the number of centroids completely included in a gold class and the number of inferred centroids. Intuitively, this type of evaluation tells us how well CLASS identifies the most tightly–bound bunches of nouns and how they correlate with sense subdivisions, as shown by the white symbols in Fig. 5. This need be complemented by other evaluation measures, as suggested in the following sections.

### 5.2.2. Noun-noun pair precision/recall

This type of evaluation is the strictest possible one. Both output lumped classes and classes in the gold standard are exploded into all possible pairs derived by coupling all and only members of the same class. A standard evaluation of the result in terms of precision–recall is then made. Precision is defined as the ratio between the number of pairs shared by the gold standard and the output partition and the number of pairs in the output partition. Recall is defined as the ratio between the number of common pairs and the number of all pairs in the gold standard (inset in Fig. 5).

This measure is extremely sensitive to the cardinality of the equivalence classes in the set partition. We can anticipate that our experiments yield a relatively high precision but a low recall. A small recall value, however significant when compared with a recall baseline of randomly selected classes, is due to a difference in "granularity" of the output partition relative to the gold standard. As already pointed out, CLASS tends to produce many smallsized equivalence classes, while the gold standard partition consists of fewer classes of medium size. To get around this bias, we introduced the following complement to pair precision/recall.

#### 5.2.3. Global analysis per class and per noun

Further information is captured through a third type of evaluation measure, by summing or averaging out precision– recall standard values (calculated on words rather than on word couples) class by class.

First we only pick up output classes built around "good" centroids, namely those centroids properly included in one of the classes in the gold standard, as wrong centroids are already penalized at the level of centroid evaluation. Then, we merge output classes whose centroids belong to the same gold class, thus forming output superclasses (OS). A standard precision-recall analysis is eventually carried out by comparing different OS's with their corresponding gold classes. First, we count how many words are contained in the intersection between each OS and its corresponding gold class. Cardinality of the intersection is then divided by the cardinality of OS (class precision) and by the cardinality of the gold class (class recall). Finally all obtained values are averaged out by summing them up and dividing the obtained sum by i) the number of elements of output classes (precision) and ii) the number of elements of gold classes (recall, black symbols in Fig. 5).

#### 5.3. Results

This section contains the results of all evaluation protocols introduced above relative to the verb–based gold standard, as summarized in Fig. 5.

CLASS shows a high level of confidence in identifying centroids (both precision and recall being around 0.8). Low pair–wise recall, on the other hand, bears witness to a considerable different degree of granularity in the way nouns are partitioned by CLASS relative to the manually– crafted gold standard. This is understandable, given the intuition–based bias of the gold standard, as opposed to the extreme sensitivity of CLASS to shifts of verb selectional preferences. It is useful to look at the way classes are automatically built around centroids: this is evaluated by measures of per-class and per-element recall, proving that centroids of relatively small-sized output classes are, in general, also centroids of medium-sized gold classes. This means that most of the elements added to an inferred centroid are consistently related to the class built around the centroid in question. On average, this is even truer for bigger gold classes than for smaller ones, per-element recall being larger than per-class recall.

All these figures combine the advantage of providing a good multivariate picture of the extent to which the system is able to replicate the gold standard with the further bonus of evaluating different phases of the classification routine in their own terms. This latter feature is invaluable for using the described evaluation protocol for purposes of progress evaluation.

Precision and recall were calculated by varying the composition of KB (*i.e.* whether it contains dictionary data only, corpus data only or both), while keeping the type of pairs (verb–object) fixed. The dictionary–only KB (circles) has precision/recall of 0.76/0.65 for centroids, 0.70/0.044 for noun–noun pairs, 0.85/0.24 for the global analysis. The corpus-only KB (squares) respectively scores 0.82/0.77, 0.76/0.24 and 0.77/0.68. Finally the mixed KB, containing both corpus and dictionary data (diamonds) scores the following figures: 0.77/0.80 (centroids), 0.73/0.20 (noun–noun pairs) and 0.77/0.68 (global per element). The best results are obtained by relying on the financial corpus only. This is reasonable since test nouns were selected among those actually attested in the financial corpus.

We also show the results obtained by projecting the same set of test nouns against corpus+dictionary KBs with different typologies of verb-noun patterns, namely: VO-KB, with verb-object patterns only; VSO-KB, with both verb-subject and verb-object patterns; and VSOP-KB, with verb-subject, verb-object, and verb-prepositional\_complement patterns. We notice that inclusion of verb-subject pairs and verbprepositional\_complement pairs in KB steadily (albeit not dramatically) improves the quality of inferred classes. We have 0.76/0.74, 0.75/0.26 and 0.76/0.64 with VO-KB (diamonds), 0.77/0.80, 0.73/0.20 and 0.77/0.68 with VSO-KB (upright triangles), and 0.79/0.84, 0.77/0.21, 0.79/0.70 (upside down triangles) with VSOP-KB.

### 6. Conclusions

Semantic classification of words is a highly context sensitive and somewhat moving target. Still, comparison of automatically-induced bottom-up classes with top-down classifications is needed to bridge the existing methodological gap between paradigmatic and syntagmatic approaches to the problem. In this paper we described a truly explorative bootstrapping method for inducing classes from distributional evidence. Moreover we introduced an evaluation protocol for projecting induced classes against a reference standard. There are at least a couple of lessons to be learned in this context.

First, hand–crafted typologies seem to be based on a considerable amount of world knowledge. This was thrown in sharp relief by the typology of words selected for in-



Figure 5: Precision/Recall for different KB's

put to CLASS, all coming from the financial domain. The financial flavour of the large majority of input words appears to have introduced a domain-specific bias into the gold standard, so that generic words were forced into the straitjacket of more specific classes. To give an example, in the golden partition of the collocates of produrre 'produce' the word sforzo 'effort' is put in the same class as accordo, dividendo, utile ('agreement', 'dividend', 'profit'), for lack of a better class, given the senses of produrre available in the reference dictionary. This sort of domain-specific pressure on the hand-crafted classification justifies the apparent paradox of getting comparatively poor results by training CLASS on dictionary data only, even if semantic classes in the gold standard were based on the sense splitting of a reference dictionary. Yet, it is important to observe that use of corpus data only does not guarantee reliable classes after lumping, even if the corresponding centroids are very good. This may suggest that stability of global precision is in fact dependent on the amount of general-purpose lexical knowledge in the KB. In a sense, domain-specific centroids are more naturally extended if more general knowledge is available. A larger variety of substitutability islands, and a resulting much tighter network of lexical associations, helps to generalize over word clusters more consistently.

While we are aware that the problem is far from being thoroughly understood, we contend that our experience suggests a principled, explorative way of tackling it.

**Acknowledgements** This work was carried out in the framework of EU project SPARKLE (LE–2111). We would like to also thank Stefano Federici for contributing to it and providing an early implementation of the clustering algorithm.

### 7. References

- Allegrini, P., Montemagni, S., Pirrelli, V. (2000) Learning Word Clusters from Data Types. To appear in *Proceedings of Coling* 2000, Saarbruecken, July 2000.
- Briscoe, T., McCarthy D., Carroll J., Allegrini P., Calzolari N., Federici S., Montemagni S., Pirrelli V., Abney S., Beil F., Carroll G., Light M., Prescher D., Riezler S., Rooth M. (1999)

Syntactic and Semantic Type and Selection. Deliverable 5.2. WP 5, EC project SPARKLE "Shallow Parsing and Knowledge Extraction for Language Engineering" (LE-2111).

- Briscoe, T., McCarthy D., Carroll J., Allegrini P., Calzolari N., Federici S., Montemagni S., Pirrelli V., Abney S., Beil F., Carroll G., Light M., Prescher D., Riezler S., Rooth M. (1999) *Acquisition System for Syntactic and Semantic Type and Selection. Deliverable 7.2.* WP 7, EC project SPARKLE "Shallow Parsing and Knowledge Extraction for Language Engineering" (LE-2111).
- Carroll, G., Light, M., Prescher, D., Rooth, M., Carroll, J., Briscoe, T., Korhonen, A., McCarthy, D., Calzolari, N., Federici, S., Montemagni, S., Pirrelli, V. (1997) *Syntactic and Semantic Type and Selection. Deliverable 5.1.* Work Package 5, EC project SPARKLE "Shallow Parsing and Knowledge Extraction for Language Engineering" (LE-2111).
- Collins Giunti Marzocco (1985) English-Italian Italian-English Dictionary. Collins Giunti Marzocco, London Firenze, 1985.
- Federici, S., Montemagni, S., Pirrelli, V. (1997) Inferring semantic similarity from Distributional Evidence: an Analogy–based Approach to Word Sense Disambiguation. In Proceedings of the ACL/EACL Workshop "Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications", Madrid, SP, July 1997.
- Federici, S., Montemagni, S., Pirrelli, V. (1999) SENSE: an Analogy–based Word Sense Disambiguation System. To appear in M. Light and M. Palmer (eds.), Special Issue of *Natural Language Engineering* on Lexical Semantic Tagging.
- Federici, S., Montemagni, S., Pirrelli, V. (1999) ROMANSEVAL: results for Italian by SENSE. To appear in M. Palmer and A. Kilgarriff (eds.), Special Issue of *Computers and the Humanities* on SENSEVAL.
- Federici, S., Montemagni, S., Pirrelli, V., Calzolari, N. (1998) Analogy-based Extraction of Lexical Knowledge from Corpora: the SPARKLE Experience. In *Proceedings of LREC–* 1998, Granada, SP, May 1998.
- Garzanti (1984) Il Nuovo Dizionario Italiano Garzanti. Garzanti, Milano, 1984.
- Miller, G.A., Charles, W.G. (1991) Contextual Correlates of Semantic Similarity. In *Language and Cognitive Processes*, 6 (1), pp. 1–28.
- Montemagni, S., Federici, S., Pirrelli, V. (1996) Resolving syntactic ambiguities with lexico–semantic patterns: an analogy– based approach. In *Proceedings of COLING–96*, Copenhagen, August 1996, pp. 376–381.
- Montemagni, S., Pirrelli, V. (1998) Augmenting WordNet–like lexical resources with distributional evidence. An application– oriented perspective. In *Proceedings of the COLING–ACL* '98 *Workshop on "Usage of WordNet in Natural Language Processing Systems*", Montreal, Canada, August 1998.
- Pereira, F., Tishby, N. (1992) Distributional Similarity, Phase Transitions and Hierarchical Clustering. In Working Notes, Fall Symposium Series. AAAI, pp. 54–64.
- Pereira, F., Tishby, N., Lee, L. (1993) Distributional Clustering Of English Words. In Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, pp. 183–190.
- Rooth, M. (Ms) Two-dimensional clusters in grammatical relations. In Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity and Generativity, AAAI 1995 Spring Symposium Series, Stanford University.
- Rooth, M., Riezler, S., Prescher, D., Carroll, G., Beil, F. (1999) Inducing a Semanticallt Annotated Lexicon via EM-Based Clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, USA, June 1999, pp. 104–111.