Learning Word Clusters from Data Types

Paolo Allegrini, Simonetta Montemagni, Vito Pirrelli
Istituto di Linguistica Computazionale – CNR
Via della Faggiola 32, Pisa, Italy
{allegrip,simo,vito}@ilc.pi.cnr.it

Abstract

The paper illustrates a linguistic knowledge acquisition model making use of data types, infinite memory, and an inferential mechanism for inducing new information from known data. The model is compared with standard stochastic methods applied to data tokens, and tested on a task of lexico–semantic classification.

1 Introduction and Background

Of late, considerable interest has been raised by the use of *local syntactic contexts* to automatically induce lexico-semantic classes from parsed corpora (Pereira and Tishby 1992; Pereira *et al.* 1993; Rooth 1995; Rooth *et al.* 1999). This family of approaches takes a pair of words (usually a verb plus a noun), and a syntactic relation holding between the two in context (usually the object), and calculates its token distribution in a training corpus. These counts define the range of more or less typical syntactic collocates selected by a verb. The semantic similarity between words is then defined in terms of substitutability in local contexts (see also Grefenstette 1994; Lin 1998): two verbs are semantically close if they typically share the same range of collocates; conversely, two nouns are semantically close if they take part in the same type of selection dependencies, *i.e.* if they are selected by the same verbs, with the same function. From this perspective, a syntactically asymmetric relation (a dependency) is reinterpreted as a semantic *co–selection*, where each term of the relation can be defined with respect to the other.

This symmetric similarity metric is often accompanied by the non trivial assumption that *the semantic classification of both verbs and nouns be symmetric* too. This is enforced by

maximizing $\prod p(v_i, n_j)$, with

$$p(v_i, n_j) \simeq \sum_{t=1}^{\omega} p(C_t)p(v_i|C_t)p(n_j|C_t), \qquad (1)$$

where $p(C_k)$ is the probability of class C_k being found in the training corpus, and $p(v_i|C_k)$ and $p(n_j|C_k)$ define the probability that verb v_i and noun n_j be associated with the semantic dimension (or *meaning component*) of class C_k. Intuitively, the *joint distribution* of functionally annotated verb–noun pairs is accounted for by assuming that each pair member independently correlates with the same *semantic dimension*, or *selection type* (Rooth 1995), a conceptual pair *defining* all pairs in the class: *e.g.* "scalar motion", "communicative action" etc.

The approach has the potential of dealing with polysemous words: as the same word can in principle belong to more than one class, there are good reasons to expect that the corresponding selection type positively correlates with one and only one sense of polysemous words. A further bonus of the approach is that it makes it explicit the *perspectivizing factor* underlying the discovered similarity of words in a class.

On a less positive side, poorly selective verbs (*i.e.* verbs which potentially combine with any noun) such as *give*, *find* or *get* tend to stick together in highly probable classes, but appear to stake out rather uninformative semantic dimensions, relating a motley collection of nouns, such as *part, way, reason* and *problem* (Rooth 1995), whose only commonality is the property of being freely interchangeable in the context of the above–mentioned verbs.

Another related issue is how many such dimensions are necessary to account for the entire variety of senses attested in the training corpus. This is an empirical question, but we contend

an important one, as the usability of the resulting classes heavily depends on it. It is common knowledge that verbs can be exceedingly choosy in the way they select their collocates. Hence, one is allowed to use the class C_k to make predictions about the set of collocates of a verb v_i, only if $P(v_i|C_k)$ is sufficiently high. Conversely, if C_k happens to poorly correlate with any verb, the set of nouns in C_k is unlikely to reflect *any* lexical selection. This compounds with the problem that the meaning of a verb v_i can significantly involve more than one semantic dimension: at the present stage of research in computational lexical semantics, no scholar has shown what function relates the meaning components of v_i to its selectional behaviour. There is wide room for further research in this area, but truly explorative tools are still needed.

Finally, the described method is acutely prone to the problem of sparse data. Although $p(C|n)$ is rightly expected to converge faster than $p(v|n)$, still convergence of $p(C|n)$ can be exceedingly slow with low frequency nouns. It is moot that sieving more and more corpus data is a solution in all cases, as word frequency is highly sensitive to changes in text genre, topic and domain (Schütze and Pedersen 1993).

2 The approach

Here we illustrate a different approach to acquiring lexico–semantic classes from syntactically local contexts. Like the family of stochastic methods of section 1, we make use of a similarity metric based on substitutability in (verb,noun,function) triples. We also share the assumption that lexico–semantic classes are inherently multidimensional, as they heavily depend on existence of a perspectivizing factor.

Yet, we depart from other assumptions. Classification of verbs and nouns is *asymmetric*: two nouns are similar if they collocate with as many *semantically diverse* verbs as possible in as many different syntactic contexts as possible. The converse applies to verbs. In other words, semantic similarity of nouns is not conditional on the similarity of their accompanying verbs, and *viceversa*. In a sense, classification *breaks the symmetry*: 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 exemplified above by the case of poorly selective verbs.

Secondly, we assume (following Fodor 1998) that the number of perspectivizing factors governing lexical selection may have the order of magnitude of the lexicon itself. The use of global semantic dimensions may smooth out lexical preferences. This is hardly what we need to semantically annotate lexical preferences. A more conservative approach to the problem, inducing *local* semantic classes, can combine applicability to real language–processing problems with the further bonus of exploring a relatively uncharted territory.

Thirdly, $p(v_i, n_j)$ appears to be too sensitive to changes in text genre, topic and domain to be expected to converge reliably. We prefer to ground a similarity metric on measuring the correlation among *verb–noun types* rather than tokens, for two basic reasons: i) verb–noun types are discrete and less prone to random variation in a (parsed) corpus, ii) verb–noun types can reliably be acquired from highly informative but hardly redundant knowledge sources such as lexica and encyclopaedias.

Finally, our information unit for measuring word similarity is *not* a couple of context–sharing pairs (*e.g.* (*set,standard*,obj) and (*set,record*,obj)) but a *quadruple* of such contexts, formed by combining two verbs with two nouns (*e.g.* (*set,standard*,obj), (*set,record*,obj), (*miss,standard*,obj) and (*miss,record*,obj)), such that they enter an *analogical proportion*.

2.1 The analogical proportion

In the present context, an analogical proportion (hereafter AP) is 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 (2) holds:

$$
\begin{aligned}
(v_k, n_i, f_m) &: (v_k, n_j, f_m) = \\
(v_t, n_i, f_n) &: (v_t, n_j, f_n),
\end{aligned} \tag{2}
$$

where terms along the two diagonals can swap place in the proportion, and identity of subscript indicates identity of values. Three aspects of (2) are worth emphasizing in this context. First, it does not require that the same syntactic function hold between all pairs, but only that functions be pairwise identical. Moreover, (2) does not cover all possible syntactic contexts where n_i, n_j, v_k and v_t may combine, *but only*

those where verb and function values co–vary.

$$(set, standard, obj) \quad : \quad (set, record, obj) =$$
$$(meet, standard, obj) \quad : \quad (meet, record, x) \quad (3)$$

We call this constraint the "same–verb–same–function" principle. As we will see in section 2.3, the principle has important consequences on the sort of similarity induced by (2). Finally, if one uses subscripts as formal constraints on type identity, then any term can be derived from (2) if the values of all other terms are known. For example given the partially instantiated proportion in (3), the last term is filled in unambiguously by substituting $x = f_n = obj$.

AP is an important generalization of the inter–substitutability assumption, as it extends the assumption to cases of functionally heterogeneous verb–noun pairs. Intuitively, an *AP* says that, for two nouns to be taken as systematically similar, *one has to be ready to use them interchangeably in at least two different local contexts.* This is where the inferential and the classificatory perspectives meet.

2.2 Mathematical background

We gave reasons for defining the similarity metric as a function of verb–noun *type* correlation rather than verb–noun *token* correlation. In this section we sketch the mathematical framework underlying this assumption, to show that, for a set of verb–noun pairs with a *unique syntactic function*, *AP* is the smallest C that satisfies eq.(1).

Eq.(1) says that v_i and n_j are conditionally independent given C, meaning that their correlation only depends on the probability of their belonging to C, as formally restated in eq.(4).

$$p(n, v|C) = p(n|C)p(v|C) \quad (4)$$

In passing from token to type frequency, we assume that a projection operator simply assigns a uniform type probability to each event (pair) with a nonzero token probability in the training corpus. From a learning perspective, this corresponds to the assumption that an *infinite memory* filters out events already seen during training. The type probability $p_T(n, v)$ is defined as in eq.(5), where N_p is the number of different pairs attested in the training corpus.

$$p_T(n, v) \quad = \quad 1/N_p \text{ if the pair is attested,}$$
$$p_T(n, v) \quad = \quad 0 \text{ otherwise.} \quad (5)$$

By eq.(4), $p_T(n, v|C) \neq 0$ if and only if $p_T(n|C) \neq 0$ and $p_T(v|C) \neq 0$. This amounts to saying that all verbs in C are freely interchangeable in the context of all nouns in C, and viceversa. We will hereafter refer to C as a *substitutability island* (*SI*). *AP* can accordingly be looked at as the minimal *SI*.

The strength of correlation of nouns and verbs in each *SI* can be measured as a summation over the strength of all *AP*s where they enter. Formally, one can define a correlation score $\sigma(v, n)$ as the probability of v and n being attested in a pair. This can be derived from our definition of $p_T(v, n)$, as shown in eq.(6), by substituting $p_T(n, v) = p_T(v)p_T(n|v)$ and $p_T(n|v) = 1/\omega(v)$, where $\omega(a)$ is the type frequency of a (*i.e.* number of different attested pairs containing a).

$$\frac{1}{N_p} = p_T(n, v) = p_T(n)p_T(v)\sigma(n, v) =$$
$$\frac{\omega(n)}{N_p}\frac{\omega(v)}{N_p}\sigma(n, v) \quad (6)$$

Eq.(6), after simplification, yields the following

$$\sigma(n, v) \propto [\omega(n)\omega(v)]^{-1}. \quad (7)$$

By the same token, the correlation function $\sigma(AP)$ relative to the 4 possible pairs in AP is calculated as

$$\sigma(AP) = p_T(v_1|n_1)p_T(n_2|v_1)p_T(v_2|n_2)p_T(n_1|v_2)$$
$$\propto [\omega(n_1)\omega(v_1)\omega(n_2)\omega(v_2)]^{-1}. \quad (8)$$

Eq.(8) captures the intuition that the correlation score between verbs and nouns in AP is an inverse function of their type frequency. Nouns and verbs with high type frequency occur in many different pairs: the less selective they are, the smaller their semantic contribution to $\sigma(AP)$.

Our preference for $\sigma(AP)$ over $\sigma(v, n)$ underlies the definition of correlation score of *SI* given in eq.(9) (see also section 4).

$$\sigma(SI) \equiv \sum_{AP \in SI} \sigma(AP). \quad (9)$$

2.3 Breaking the symmetry

In section 2.2 we assumed, for the sake of simplicity, that verbs and nouns are possibly related through one syntactic function only. In a

proportion like (2), however, the syntactic function is allowed to vary. Nonetheless each related *SI* contains nouns which always combine with a given verb *with one and the same syntactic function.* Clearly, the same is not true of verbs. Suppose that an *SI* contains two verbs v_k and v_t (say *drive* and *pierce*) and two nouns n_i and n_j (say *nail* and *peg*) that are respectively object and subject of v_k and v_t. The type of similarity in the resulting noun and verb clusters is of a completely different nature: in the case of nouns, we acquire *distributionally parallel words* (e.g. *nail* and *peg*); in the case of verbs, we get *distributionally correlated words* (say *drive* and *pierce*) which are *not* interchangeable in the same context. Mixing the two types of distributional similarity in the same class makes little sense. Hereafter, we will aim at maximizing the similarity of distributionally parallel nouns. In doing so, we will use functionally heterogeneous contexts as in (2). This breaks classification symmetry, and there is no guarantee that semantically coherent verb clusters be returned.

3   The method

The section illustrates an application of the principles of section 2 to the task of clustering the set of objects of a verb on the basis of a repository of functionally annotated contexts.

3.1   The knowledge base

The training evidence is a Knowledge Base (*KB*) of functionally annotated verb–noun pairs, instantiating a wide range of syntactic relations:

a) verb–object, e.g. (*causare, problema,* obj) 'cause-problem';

b) verb–subject, e.g. (*capitare, problema,* subj) 'occur-problem';

c) verb–prepositional_complement, e.g. (*incappare, problema, in*) 'run_into-problem'.

The *KB* contains 43,000 pair types, automatically extracted from different knowledge sources: dictionaries, both bilingual and monolingual (Montemagni 1995), and a corpus of financial newspapers (Federici *et al.* 1998). The two sources reflect two different modes of lexical usage: dictionaries give typical examples of use of a word, and running corpora attest actual usage of words in specific embedding domains. These differences have an impact on the typology of senses which the two sources

provide evidence for. General dictionaries testify all possible senses of a given word; 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.

3.2   The input words

There is abundant psycholinguistic evidence that semantic similarity between words is eminently context–sensitive (Miller and Charles 1991). Moreover, in many language–processing tasks, word similarity is typically judged relative to an actual context, as in the cases of syntactic disambiguation (both structural and functional), word sense disambiguation, and selection of the contextually appropriate translation equivalent of a word given its neighbouring words. Finally, close examination of real data shows that different word senses select classes of complements according to different dimensions of semantic similarity. This is so pervasive, that it soon becomes impossible to provide an effective account of these dimensions independently of the sense in question.

Evaluation of both accuracy and usability of any automatic classification of words into semantic clusters cannot but artificially elude the basic question "similar in what respect?". Our choice of input words reflects these concerns. We automatically clustered the set of objects of a given verb, as they are attested in a test corpus. This yields *local* lexico–semantic classes, *i.e.* conditional on the selected verb head, as opposed to global classes, *i.e.* built once and for all to account for the collocates of *any* verb.

Among the practical advantages of local classification 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.

3.3   Identification and ranking of noun clusters

For the sake of concreteness, let us consider the following object–collocates of the Italian verb

```
{APPESANTIMENTO, CRESCITA,
     FLESSIONE, RIALZO}      :    CAUSARE/O, REGISTRARE/O
    {CRESCITA, FLESSIONE}    :    CAUSARE/O, EVIDENZIARE/S,
                                  MEDIARE/O, MOSTRARE/O,
                                  PRESENTARE/S, PRESENTIRE/O
                                  REGISTRARE/O, REGISTRARE/S
        {CRESCITA, GUAIO}    :    CAUSARE/O, PROVOCARE/O
     {CRESCITA, PROBLEMA}    :    AVERE/S, CAUSARE/O, EVIDENZIARE/O
                                  PORRE/S, PRESENTARE/S
      {CRESCITA, RITARDO}    :    CAUSARE/O, USARE/S
    {FLESSIONE, PROBLEMA}    :    CAUSARE/O, PRESENTARE/S, STARE/S
      {FLESSIONE, RIALZO}    :    CAUSARE/O, REGISTRARE/O
                                  REGISTRARE/S, SUBIRE/O
        {GUAIO, PROBLEMA}    :    CAUSARE/O, CAVARE − SI/S, INCAPPARE/S
{RIDIMENSIONAMENTO, RITARDO} :    CAUSARE/O, GIUSTIFICARE/O
                             ...
```

Figure 1: Some SIs relative to the collocates of the headword *causare*.

causare 'cause', as they are found in a test corpus:

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

Clustering these input words requires preliminary identification of Substitutability Islands (SIs). An example of SI is the quadruple formed by the verb pair *causare* 'cause' and *incappare* 'run into' and the noun pair *guaio* 'trouble' and *problema* 'problem', where members of the same pair are inter–substitutable in context, given the constraints enforced by the AP type in (2). Note that *guaio* and *problema* are objects of *causare*, and prepositional complements (headed by *in* 'in') of *incappare*. This makes it possible to maximize the similarity of trouble and problem across functionally heterogeneous contexts.

Bigger SIs than the one just shown will form as many APs as there are quadruples of contextually interchangeable nouns and verbs. We consider a lexico–semantic cluster of nouns the projection of an SI onto the set of nouns. Fig.1 illustrates a sample of noun clusters (between curly brackets) projected from a set of SIs, together with a list of the verbs found in the same SIs (the suffix 'S' stands for subject, and 'O' for object). Due to the asymmetry of classification, verbs in SIs are not taken to form part of a lexico–semantic cluster in the same sense as nouns are.

```
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{FLESSIONE,PROBLEMA}
```

Figure 2: Nine top-most scored noun clusters

Not all projected noun clusters exhibit the same degree of semantic coherence. Intuitively, the cluster {*appesantimento crescita flessione rialzo*} 'increase in weight, growth, decrease, rise' is semantically more appealing than the cluster {*crescita problema*} 'growth problem' (Fig.1).

A quantitative measure of the semantic cohesion of a noun cluster C_N is given by the correlation score $\sigma(SI)$ of the SI of which C_N is a projection. In Fig.2 noun clusters are ranked by decreasing values of $\sigma(SI)$, calculated according to eq.(9).

3.4 Centroid identification

Noun clusters of Figs.1 and 2 are admittedly considerably fine grained. A coarser grain can be attained trivially through set union of intersecting clusters. In fact, what we want to obtain is 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.

In the algorithm evaluated here this is achieved in two steps: i) first, we select the best possible *centroids* of the prospective classes among the noun clusters of Fig.2; secondly, ii) we lump outstanding clusters (*i.e.* clusters which have not been selected in step i)) around the identified centroids. In what follows, we will only focus on step i). Results and evaluation of step ii) are reported in (Allegrini *et al.* 2000).

In step i) we assume that centroids are disjunctively defined, maximally coherent classes; hence, there exists no pair of intersecting centroids. The best possible selection of centroids will include non–intersecting clusters with the highest possible cumulative score. In practice, the best centroid corresponds to the

cluster with the topmost $\sigma(SI)$. The second best centroid is the cluster with the second highest $\sigma(SI)$ and no intersection with the first centroid, and so on (the i-th centroid is the i-th highest cluster with no intersection with the first $i-1$ centroids) until all clusters in the rank are used up. Clusters selected as centroids in the *causare* example above are: {GUAIO PROBLEMA}, {RIDIMENSIONAMENTO RITARDO}, {CRESCITA FLESSIONE}.

Clearly, this is not the only possible strategy for centroid selection, but certainly a suitable one given our assumptions and goals. To sum up, the targeted classification is *local*, *i.e.* conditional on a specific verb head, and *orthogonal*, *i.e.* it aims at identifying maximally disjunctive classes with high correlation with the principal meaning components of the verb head. This strategy leads to identification of the different senses, or possibly meaning facets, of a verb. In their turn, noun clusters may capture subtle semantic distinctions. For instance, a distinction is made between incremental events or results of incremental events, which presuppose a scalar dimension (as in the case of {*crescita flessione*} 'growth, decrease') and rescheduling events, where a change occurs with respect to a previously planned event or object (see the centroid {*ridimensionamento ritardo*} 'reduction delay').

4 Experiment and evaluation

We were able to extract *all* SI's relative to the entire KB. However, we report here an *intrinsic* evaluation of the accuracy of acquired centroids which involves only a small subset of our results, since provision of a reference class typology is extremely labour–intensive. [1]

We consider 20 Italian verbs and their object collocates.[2] The object collocates were automatically extracted from the "Italian SPARKLE Reference Corpus", a corpus of Italian financial

---

[1] For an *extrinsic* evaluation of the proposed similarity measure the reader is referred to (Montemagni *et al.* 1996; Briscoe *et al.* 1999; Federici *et al.* 1999a).

[2] The test verbs are: *aggiungere* 'add', *aiutare* 'help', *aspettare* 'expect', *cambiare* 'change', *causare* 'cause', *chiedere* 'ask', *considerare* 'consider', *dare* 'give', *decidere* 'decide', *fornire* 'provide', *muovere* 'move', *permettere* 'allow', *portare* 'bring', *produrre* 'produce', *scegliere* 'choose', *sentire* 'feel', *stabilire* 'establish', *tagliare* 'cut', *terminare* 'end', *trovare* 'find'.

newspapers of about one million word tokens (Federici *et al.* 1998).

For each test verb, an independent classification of its collocates was created manually, by partitioning the collocates into disjoint sets of semantically coherent lexical preferences, each set pointing to distinct senses of the test verb, according to a reference monolingual dictionary (Garzanti 1984). This considerably reduces the amount of subjectivity inevitably involved in the creation of a reference partition, and minimizes the probability that more than one sense of a polysemous noun can appear in the same class of collocates.

The inferred centroids, selected from clusters ranked by $\sigma(SI)$ defined as in (9), are projected against the reference classification. Precision is defined as the ratio between the number of centroids properly included in one reference class and the number of inferred centroids. Recall is defined as the ratio between the number of reference classes which properly include at least one centroid and the number of all reference classes.

Fig.3 shows results for the sets of object collocates of polysemous test verbs only, as monosemous verbs trivially yield 100% precision–recall. An average value over the sets of object collocates of all verbs is also shown, with 86%–88% of precision–recall. Another average value is also plotted (as a black upright triangle), obtained by ranking noun clusters by $\sigma(SI)$ calculated as in (10). This average value (53%–53% precision–recall) provides a sort of baseline of the difficulty of the task, and sheds considerable light on the use of APs, rather than simple verb–noun pairs, as information units for measuring internal cohesion of centroids.

$$\sigma(SI) \equiv \sum_{(n,v)\in SI} \sigma(n,v) \qquad (10)$$

5 Conclusion

We described a linguistic knowledge acquisition model and tested it on a word classification task. The main points of our proposal are:

- classification is asymmetric, grounded on principles of machine learning with infinite memory;

- the algorithm is explorative and non–reductionist; no *a priori* model of class dis-
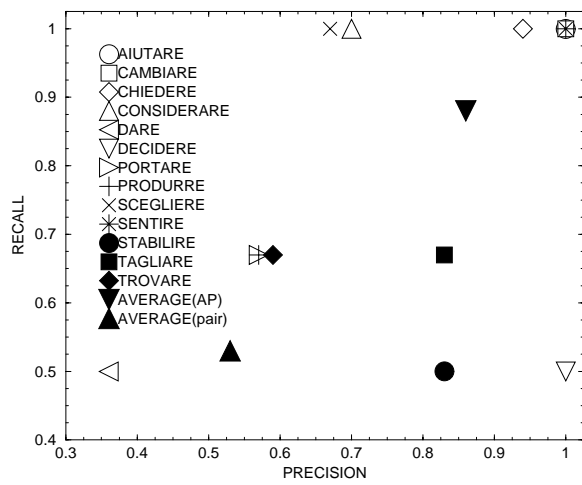
Figure 3: Centroid precision and recall for object collocates of polysemous verbs.

tribution is assumed;

- classification is modelled as the task of forming a web of context–dependent semantic associations among words;

- the approach uses a context–sensitive notion of semantic similarity;

- the approach rests on the notion of analogical proportion, which proves to be a reliable information unit for measuring semantic similarity;

- analogical proportions are harder to track down than simple pairs, and interconnected in a highly complex way; yet, reliance on data types, as opposed to token frequencies, makes the proposed method computationally tractable and resistant to data sparseness.

References

Allegrini P., Montemagni S., Pirrelli V. (2000) Controlled Bootstrapping of Lexico-semantic Classes as a Bridge between Paradigmatic and Syntagmatic Knowledge: Methodology and Evaluation. In *Preceedings of LREC-2000*, Athens, Greece May-June 2000, pp. 601-608.

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).

Federici, S., Montemagni, S., Pirrelli, V. (1999) SENSE: an Analogy–based Word Sense Disambiguation System. In M. Light and M. Palmer (eds.), Special Issue of *Natural Language Engineering* on Lexical Semantic Tagging.

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.

Fodor, J.A. (1998) *Concepts. Where Cognitive Science Went Wrong.* Clarendon Press, Oxford, 1998.

Garzanti (1984) *Il Nuovo Dizionario Italiano Garzanti.* Garzanti, Milano, 1984.

Grefenstette, G. (1994) *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers, Boston, 1994.

Lin D. (1998) Automatic Retrieval and Clustering of Similar Words. In *Proceedings of COLING-ACL'98*, Montreal, Canada, August 1998.

Miller, G.A., Charles, W.G. (1991) Contextual Correlates of Semantic Similarity. In *Language and Cognitive Processes*, 6 (1), pp. 1–28.

Montemagni, S. (1995) *Subject and Object in Italian Sentence Processing.* PhD Dissertation, UMIST, Manchester, UK, 1995.

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.

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.

Schütze, H., Pedersen, J. (1993) A vector model for syntagmatic and paradigmatic relatedness. In *Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research*, Oxford, England, 1993, pp. 104–113.