Coping with WORDNET Sense Proliferation

Alessandro Artale, Anna Goy*, Bernardo Magnini, Emanuele Pianta & Carlo Strapparava
IRST, Istituto per la Ricerca Scientifica e Tecnologica
I-38050 Povo TN, ITALY
[artale | magnini | pianta | strappa@irst.itc.it]
*Dipartimento di Informatica - University of Torino, Italy
[goy@di.unito.it]

Abstract

WORDNET makes a great number of fine-grained word sense distinctions. However, what could be seen as an advantage has often been considered a problem from a computational point of view. A great number of sense distinctions makes harder the problem of word sense disambiguation. One way to face this issue is reducing the number of senses, for example by grouping them into equivalence classes which abstract on some aspects of the meanings of words. In this paper we will try a different approach. Although we recognize that some sense distinctions in WORDNET are dubious, we prefer to keep the semantic richness of WORDNET and to make some proposals to extend it in order to make the task of word sense disambiguation easier.

Introduction

Lexical Semantic research in the last years (Calzolari, 1992; Pustejosky, 1995) has emphasized the centrality of the notion of word sense in the organization of a computational lexicon. The availability of word sense repositories, such as WORDNET (Miller, 1990), increased the interest for the realization of concrete NLP applications that can take advantage of sense distinctions.

However, a well known problem for the computational use of WORDNET is that, although it includes a large amount of word senses, just few information are available that can be used for sense disambiguation. Although some of the WORDNET sense distinctions are ill-motivated, in this paper we take the view that the large majority of them are reasonable. In this paper we make some proposals for extensions to WORDNET, which can be used to improve word sense disambiguation.

Some of the data presented in the paper are derived from Italian WORDNET (Magnini & Strapparava, 1997; Magnini *et al.*, 1994), an extension of the English WORDNET to Italian, currently under development at Irst.

The approach we propose will be concretely experimented in the context of the LE TAMIC-P project (Transparent Access to Multiple Information for the Citizen - Pensions), an information access system specifically designed for the Public Administration domain.

The paper is organized as follows. In section 1 we introduce a new semantic relation (*pertain-to-subject*), useful to disambiguate word senses against topical contexts. Section 2 suggests to extend verbal frames with more accurate selectional restrictions expressed as logical compositions of noun-synsets. Section 3 analyzes disambiguation problems of WORDNET adjective senses

1. Adding Subject Field Labels

Experimental work by Leacock, Towell & Voorhees (1995) shows that knowing the topic of the discourse (*topical context*), allows current algorithms for word disambiguation to select the correct sense of a word in 70% of cases; human subjects seem to perform the same task with comparable results. For example if a human subject is given the word *sheet* and the topical context “sleeping”, he/she is very likely to select the meaning “bed linen” instead of “piece of paper”. Miller (1995) suggests that topical context could be used to choose among WORDNET senses. For instance if the domain of discourse is limited to air travel, only one of the nine senses listed in WORDNET 1.6 for the word *flight* is likely to occur.

To use topical context for disambiguating WORDNET senses at least the following steps are needed:

1. to define a set of subject codes;
2. to associate subject codes to WORDNET synsets;
3. to label discourse segments with subject codes.

Point 3 is out the scope of this article; in the rest of this section we will concentrate on points 1 and 2.

The first issue is what counts as a subject code. Let us consider first how subject codes are used in existing lexical resources. If we look at paper dictionaries we find that the best approximation to the notion of subject code are field labels such as: `Anat` (anatomy), `Archeol` (archeology), `Bot` (botany), etc. The number of such labels varies among dictionaries. Here is a sample list of seven dictionaries of different nature; five of them are monolingual, two bilingual, four are large size, one is medium and two are pocket-size; relevant languages are English, Italian, Spanish and Swedish. We give the number of distinct field labels for each dictionary:

- *Oxford Adv. Learner's Dict. of Current Eng.* (English monolingual, mid-size): 58
- *DISC* (Italian monolingual, large-size): 41
- *Garzanti* (Italian monolingual, large-size): 58
- *Palazzi* (Italian monolingual, large-size): 129
- *Goeteborg Lexical DataBase* (Swedish monolingual, large-size): 95
- *Herder* (Bilingual Italian-Spanish, pocket-size): 48
- Collins GEM (Bilingual Ita-Eng, pocket-size): 47

The union set of all the labels includes 177. Here is a table reporting how many labels occur in how many dictionaries

labels	dictionaries
17	7
23	6
6	5
12	4
20	3
29	2
70	1

As the date shows, a relatively small set of labels (40) occurs in almost all dictionaries whereas a large set (99) occurs only 1 or 2 times.

It is worth noting that some of WordNet glosses include a field label of this type (at the beginning of the definition between parenthesis). See for example the synsets for *computer program* and its hypernyms.

```
{program, programme, computer program, ...}
-- ((computer science) a sequence
   of instructions ...)
   => {software, software system}
    -- ((computer science) written programs
       or procedures ...)
      => {code, computer code
         -- ((computer science) the symbolic
          arrangement of data ...)
```

WordNet 1.6 glosses include approx. 200 subject labels. The use of labels is quite free and there seems not to be an established set of labels that all lexicographers use. For example as label of medical terms one can find one of the following codes: med and pathology, med, medical, medicine and pathology. Some labels are very idiosyncratic, i.e. they label only one synset (for instance: bacteriology, classical antiquity, matrix algebra). Approx. 3500 synsets are labeled by subject codes, i.e. 3.5% of all synsets. We can conclude that the use of subject label in WordNet 1.6 is not systematic and has a quite limited coverage. Just to make an example, no field label distinguishes the two senses of *mouse*:

```
{mouse} -- (any of numerous small rodents ...)
{mouse} -- (a hand-operated data input ...)
```

The question now becomes: are field labels, as they are used by paper dictionaries and WordNet definitions, suitable for word sense disambiguation? The answer is: probably no. Field labels are manly used to signal the specialistic use of a word, words that are used in a specific discipline, craft or activity (Landau, 1994). They are not used to disambiguate the meaning of words. Two are the consequences: (a) many ambiguous words don't have any field label (because they don't belong to any specialistic terminology); (b) only a very restricted number of labels refer to non specialistic subjects. To overcome these shortcomings let us try a different approach. If we look at the subject labels used by dictionaries we see that most of them are words that we can look up in WordNet. Thus, we could use the synsets themselves as subject identifiers. Then, to associate word senses to subject fields we need to introduce a new semantic relation between synsets. We call it the *pertain-to-subject* relation. Its meaning is as follows:

the lexical concept identified by synset S1 pertains to the subject field identified by synset S2

This solution has at least two main advantages: 1) the definition of subject codes reduces to a selection between the existing WordNet synsets; 2) we can associate subject fields to synsets by introducing an instance of a known class of WordNet relations (semantical relations)

We tried to map the kind of subject labels used by paper dictionary onto WN synsets, and found that it is always possible to find a one to one correspondence. This is quite a good result as one is often willing to couple information coming from WordNet with the information coming from the definitions and the glosses available through Machine Readable Dictionaries. For example this is a crucial task in the project for the semi-automatic construction of the Italian version of WordNet, undergoing at IRST. Note that, contrary to what happens with paper dictionaries, it is sometimes difficult to match the subject labels used in WN glosses with WN synsets: for instance it is not possible to find a synset corresponding to the subject Scandinavian folklore. That means that sometimes WN lexicographers use subject descriptions which are more specific than any existing lexical concept in WordNet. Generally speaking, finding the right level of granularity for topical contexts is a problematic issue. We feel that using the kind of granularity supplied by WordNet synsets is a sensible and balanced solution to the problem. So in the above example we would use a less specific subject, that is the concept identified by the {folklore} synset.

So far we proposed a solution for step 1 (defining a set subject codes). The second step (adding pertain-to-subject relations) need to be done by hand. An experimental project at this end is undergoing at IRST. Notice that the pertain-to-subject relation has an interesting feature that makes our task easier. Actually, we can assume that if S1 pertains-to-subject S2, then the same relation holds for all the hyponyms of S1. Thus, we can use WordNet hierarchy to add subject field information in a very compact way.

2. Adding Verbal Selectional Restrictions

Selectional restrictions provide an explicit semantic information that the verb supplies about its arguments (Jackendoff, 1990). Although this information could be profitably used for verbal sense disambiguation, there seems to be at least two open questions relevant for the introduction of selectional restrictions into the WordNet framework: (i) a decision has to be taken whether a selectional restriction is a lexical relation, i.e., it has to be associated to a word, or it is a conceptual one, i.e., it has to be associated to a synset; (ii) it is necessary to individuate the appropriate degree of details in the description of selectional restrictions.

As far as the first point is concerned, currently WordNet implements selectional restrictions as lexical relations, that is, syntactic frames and their restrictions are associated to verbal word forms. This is necessary because verbs in the same synset can have different superficial behaviors and so they need different selectional restrictions. In the following example the Italian verbs “scrivere”(*write*) and “redigere” (*indite*), which are synonyms in the synset Write-Compose (see Figure 1), admit different selectional restrictions:

(1) Proust ha scritto/composto/*redatto la “Recherche” nel 1912. (*Proust wrote/composed/*indited the “Recherche” in 1912*)

However, it seems also reasonable that verbs belonging to the same synset share common properties (because they are synonyms) and that these properties can be represented at the synset level. In our view a verbal synset is an homogeneous conceptual representation of a state/action which is linguistically lexicalized by the verbs belonging to the synset. As such, a verbal synset can be described by a fixed number of participants to the state/action, each of them playing a semantic role and each of them restricted to be of a particular kind. For instance, the Write-Compose synset require an agent, who has to be a human, and a theme, that has to be a kind of written stuff.

Given the above considerations we propose to represent selectional restrictions at the synset level where they provide generic and typical restrictions over semantic participants to the state/action described by a verbal synset. As far as more specific uses of a single verb form are concerned (as it is the case for the verb *indite* in sentence 1) more peculiar information need to be added to the single verb entry. This latter point will not be further developed here.

Synset Label	**Italian Synset**	**English Synset**
Write-Compose	{scrivere redigere comporre}	{write compose pen indite}
Write-Trace	{scrivere tracciare}	{write trace}
Write-Music	{scrivere comporre}	{compose write write_music}
Write-Communicate	{scrivere comunicare_per_iscritto}	{write communicate_by_writing}
Write-Publish	{scrivere pubblicare}	{publish write}
Write-Send	{inviare mandare scrivere spedire}	{mail write post send}

Figure 1. Correspondences between Italian and English synsets for the verb ‘*scrivere’* (write)

As far as the level of description of selectional restrictions is concerned, all the English verbs of WordNet are described resorting to a set of 35 different syntactic frames, which in turn include only two restrictions, that is “Something” and “Somebody”. For example, the frames provided for the verb “Write” in the synset {publish, write} are given in the form of two patterns, where the dots can be substituted by the verb stem:

Something ...s
Somebody ...s Something

This level of description in many cases results to be too general for a word sense disambiguation task. In (Artale *et al.* 1997) we argued that a more detailed level of selectional restrictions than the one implemented in WordNet would make sense disambiguation more effective. In particular we suggested to define selectional restrictions as a logical combination of WordNet noun synset. The appropriate combination of synsets for an argumental position has to be both enough general to preserve all the human readings, and enough restricted for discriminating among different senses of both verb and noun. Figure 2 shows selectional restrictions for the senses of the verb *write*. For each sense a conventional name which unambiguously identify the synset is reported, as well as the argumental positions admitted for that sense, along with the indication of the selectional restrictions.

We approached the problem of selecting the right verb sense by finding the appropriate selectional restrictions. This revealed as a difficult and time consuming task. In order to achieve a good trade-off between discrimination power and precision level we adopted an empirical process with successive steps of refinement. We started with general selectional restrictions and then we validate them against a previously collected corpus. But it is also true that some form of reusability apply, at least when building selectional restrictions for the various senses of the same verb. Let us consider the *write* senses. The restriction for the Object of Write-Communicate is just the union of the ones we imposed for the Write-Compose and Write-Trace. We built the Object restriction for the Write-Send sense by refining the Object restrictions of Write-Compose and Write-Communicate senses by looking for all kinds of Communications that we can send. A simple look at the selectional restrictions shows an evidence for a hierarchical relation between the two senses Write-Send and Write-Communicate, also confirmed empirically. We would note that, every time a *troponymy* relation between two verbs holds - defined as the co-occurrence of both lexical implication and temporal co-extension between two verbs - a subsumption relation between the correspondent selectional restrictions holds, too. Obviously, a hierarchical structure would make easier the addition of new selectional restrictions

An experiment was made that shows both the plausibility of WordNet senses for describing lexical entries and the usability of WordNet for carrying out lexical discrimination. In the experiment a small number of lexical entries was built to allow an Italian parser to analyze a set of sentences. Whenever the parser. tries to build a (partially recognized) constituent it incrementally verifies the admissibility of the semantic part of such a constituent. In particular, whenever a noun is associated with a verbal argument an Isa function is triggered to check whether the synset of the noun is subsumed by the selectional restriction of the corresponding verbal argument. As soon as this semantic test fails the constituent is rejected.

As an example of use of selectional restrictions for disambiguation, consider the following sentence: *Peter writes its name to Mary*, where *name* is subsumed by the synset Signal. The only allowed senses for *write* are Write-Trace and Write-Communicate. Indeed, since a Signal cannot be the object of a composition the sense Write-Compose is discarded. This same argument applies to the remaining senses. It is interesting to note that, even if this is an ambiguous case, the

WordNet Synset	**Subject**	**Object**	**Indirect-Object**
Write-Compose	Somebody	Communication ∧ ¬Signal	--
Write-Trace	Person	Signal ∨ Measure-Amount ∨ Language-Unit ∨ Property	--
Write-Music	Person	Sheet-Music	--
Write-Communicate	Somebody	(Communication∧ ¬Signal) ∨ (Signal ∨ Measure-Amount ∨ Language-Unit ∨ Property)	Somebody
Write-Publish	Somebody	Written-Material	Print-Media ∨ Publishing-House
Write-Send	Somebody	Correspondence ∨ Missive ∨ Message	Somebody

Figure 2. Synset Selectional Restrictions

Experimental Setting	**# of readings**	**Discrimination Rate**	**Precision**
Without discrimination	1201	0%	10%
Discrimination with WordNet Frames	688	43%	18%
Discrimination with WordNet Full Hierarchy	164	86%	74%
Human Judgment	122	90%	100%

Figure 3. Quantitative results obtained on 60 sentences

preferred reading is the one of Write-Communicate since the noun phrase *to Mary* fills the indirect object required for this sense.

Two hypotheses on selectional restrictions have been checked, i.e., the one with general WordNet frames and the other with more refined selectional restrictions. The analyses produced by a parser have been compared with the set of interpretations given by a human. Results are reported in Figure 3.

These results have to be interpreted considering that the focus of the experiment is on selectional restrictions, which of course is just one among the various kinds of information occurring during lexical discrimination. It is worth mentioning here some other crucial information sources: (i) world knowledge (e.g., it is very strange to Write a Paper on a Newspaper-Periodic); (ii) aspectual properties of the verb, e.g., it is very difficult to interpret the sentence "*Mary is writing an article on the newspaper*" with the Write-Publish sense, because publishing is a culminative process. For what concerns the first point, a WordNet sense should provide information about the sense related verbal default arguments (Pustejosky, 1995). This is relevant because sense disambiguation is crucially affected by the kind of adjuncts the sense admits (Gomez *et al.* 1997). Consider the following sentences:

(2) Mary wrote a letter
(3) Mary wrote a letter on the blackboard.

While in sentence (2) *write* is ambiguous between Write-Compose and Write-Trace, the verbal adjuncts *on the blackboard* in sentence (3) eliminates the Write-Compose sense allowing only the Write-Trace interpretation. This kind of disambiguation can be carried on by adding more structure to a verb synset. As proposed in (Gomez *et al.* 1997), we can associate to each verb a frame-like representation where every thematic role is annotated with the syntactic relation introducing it - including the possible preposition allowed - together with the semantic restriction required by the thematic role. In this work hypothesis, the verb hierarchy would be crucial since we could exploit the inheritance mechanism during the insertion of new items.

3 Adjective Polysemy

One aspect of the word disambiguation task, when interpreting a sentence, is related to head-modifier constructions. In such constructions, the disambiguation usually consists in choosing the proper sense for the modifier, given the one of the head[1]. Among head-modifier constructions, noun-adjective ones are particularly interesting since the meaning of adjectives strongly depends on the context, and the main feature of the linguistic context is the noun they modify.

One of the best known examples of the difficulty in selecting the proper sense for the modifier is the adjective *good*, when modifying different nouns (*good news*, *good knife*, *good sandwich*, *good wife*, etc.): WordNet lists 25 sense for *good* (as an adjective). A simpler example are adjectives which denote psychological states (*sad*, *happy*, etc.)[2]. Let's consider the Italian adjective *allegro* (*happy*/*cheerful*). The Italian dictionary Palazzi-Folena gives the following definition for *allegro*:

[1] This task relies on the assumption that the head has already been disambiguated; actually, these two steps, i.e. the choice of the proper sense for the head and then for the modifier, need not to be sequential.

[2] For an analysis of such adjectives, see (Goy 1998).

allegro **1.** che sente o dimostra allegria (stato d'animo lieto e festoso, allegrezza); di temperamento o disposizione allegra - *è un tipo allegro*. **2.** brioso, che infonde allegria - *colore, spettacolo allegro, musica allegra*.[3]

Adjectival entries in Italian WORDNET are still under development; however we assume that the synsets available for *allegro* will correspond to these two meanings:

Synset Label	**Italian Synset**
`Allegro-stative`	`{allegro contento festoso gaio lieto}`
`Allegro-causative`	`{caldo vivace piacevole divertente}`

Figure 4. Italian synsets for the adjective *allegro* (happy/cheerful)

Let's consider two sentences:

(4) Papà è allegro questa sera (*Dad is happy tonigth*)
(5) Vorrei comprarmi un quadro allegro per il soggiorno
(*I would like to buy a cheerful painting for the living room*)

In (4) *allegro* refers directly to the psychological state of a human being (the one denoted by "papà"), while in (5) its meaning is something like "which cause happiness/cheerfulness in people watching it". The main point here is that we can disambiguate *allegro*, i.e. we can select the "causative" sense, only by taking into account the semantic properties of the noun it refers to, i.e. *quadro* (*painting*), which denotes an artifact.
As far as psychological adjectives are concerned, we can have one more reading, i.e. the "manifestative" one (see Bouillon 1996), as in (6), where *affettuosa* (*loving/affectionate*) means "that expresses/manifests love".

(6) Maria mi ha scritto una lettera molto affettuosa
(*Maria wrote me a very affectionate letter*)

The availability of these three interpretations - "stative", as in (4), "causative", as in (5), and "manifestative", as in (6) - depends on the kind of adjective involved, since not every psychological adjectives allow all three, but also on the semantic type of the modified noun.
As far as the first information is concerned, the availability of one, two, or three readings in encoded in the number of senses of the adjectival entry. As for the interaction with the meaning of the noun, intuitively, the disambiguation strategy is the following:

- if the noun denotes an `event` (*scampagnata allegra - cheerful trip*), then the "stative" reading is not available;
- if the noun denotes a `physical object`, then we need a distinction between (at least) artifacts and natural kinds:
  - – if it is an `artifact` (*quadro allegro - cheerful painting*), the expression seems to be ambiguous (at least) between two readings: the "causative" ("a painting that makes people watching it cheerful") and the "manifestative" ("a painting that expresses the painter's cheerfulness");
  - – if it is a `natural kind` (*fiore allegro - cheerful flower*), then only the "causative" reading seems to be available ("a flower that makes people watching it cheerful")
- if the noun denotes a `human being`, then we have different possibilities:
  - – if it refers to a "role" (*pittore allegro - cheerful painter*), then all three reading are available[4] ("a cheerful person, who is a painter", "a painter whose paintings make people watching it cheerful", "a painter whose paintings expresses his/her cheerfulness");
  - – if it does not refers to any "role" (*ragazzo allegro - cheerful boy*), then the "stative" reading seems to be strongly preferred.

If each sense in the WORDNET entry for the adjective contains the selectional restriction for the argument to be modified, then the disambiguation task could be performed by matching such restrictions with the semantic type of the head noun, i.e. with its WORDNET synset (or one of its hyperonyms). For instance, the hyperonym hierarchy of the synset corresponding to the first sense of *dad* (*papa*) contains

```
=> person, individual, human, ...;
```

the hyperonym hierarchy of the synset corresponding to the first sense of *painting* (*quadro*) contains

```
=> artifact, artefact.
```

On the adjective side, the `Allegro-stative` synset will have the selectional restriction `human`, while the `Allegro-causative` one will have `artifact`: this information is the one that enable the linguistic interpreter to choose the proper sense in cases as (4) and (5).

Conclusions

In this paper we made three proposals for coping with the so called problem of sense proliferation in WordNet. Instead of reducing the richness of WordNet sense distinctions, we propose to add new information useful for the sense disambiguation task.

Acknowledgements

This work has been partially founded by the LE-4253 TAMIC-P project. Italian WORDNET has been partially developed in the framework of the ILEX (Italian Lexicon) project.

Bibliographical References

Artale, A, Magnini, B., Strapparava, C., (1997). WordNet for Italian and Its Use for Lexical Discrimination. In Maurizio Lenzerini (Ed.) AI*IA 97: Advances in Artificial Intelligence. Proceeedings of the 5th Congress of the Italian Association for Artificial Intelligence, Roma, Italy, 16-19 settembre 1997, Springer Verlag.

Briscoe, T. (1991). Lexical Issues in Natural Language Processing. In Klein E. and Veltman F. (eds.): *Esprit*

[3] **1.** Who feels or shows happiness (cheerful mood); with happy temperamento or disposition - *he is a happy guy*. **2.** brioso, that infuses happiness - *cheerful color, show, music*.

[4] Maybe with different degrees of acceptability.

Symposium on Natural Language and Speech, Berlin, Springer-Verlag.

Bouillon, P. (1996). Mental states adjectives: the perspective of generative lexicon. In *Proceedings of COLING-96*, Copenhagen.

Calzolari N. (1992). Acquiring and Representing Semantic Information in a Lexical Knowledge Base. In Pustejovsky, J. & Bergler, S. (eds.) *Lexical Semantics and Knowledge Representation*, Springer-Verlag, Berkeley.

Delmonte, R., Ferrari, G., Goy, A., Lesmo, L., Magnini, B., Pianta, E., Stock, O., Strapparava, C. (1996). ILEX - Un dizionario computazionale dell'Italiano. In *Proceedings of the 5th Convegno Nazionale della Associazione Italiana per l'Intelligenza Artificiale*, Napoli, 26-28 settembre 1996.

Gomez, F., Segami, C. & Hull, R. (1997). Determing Prepositional Attachment, Prepositional Meaning, Verb Meaning, and Thematic Roles. *Computational Intelligence* vol. 13, num. 1.

Goy A. (1998) Il ruolo della semantica lessicale nella comprensione del linguaggio naturale: il caso degli aggettivi in italiano, PhD thesis, Università di Torino.

Jackendoff, R. (1990). *Semantic Structures*. Current Studies in Linguistics. The MIT Press, Cambridge, Massachusetts/London, England.

Landau, S.I. (1994). *Dictionaries: The art & craft of lexicography*. New York: The Scribner Press.

Leacock, C., Towell G. & Voorhees E.M. (1996). Towards building contextual representations of word senses using statistical models. In Boguraev, B. & Pustejovsky, J. (Eds.), *Corpus processing for lexical acquisition* (pp. 97—113). Cambridge, MA: The MIT Press.

Magnini, B. & Strapparava, C. (1997). Costruzione di una base di conoscenza lessicale per l'italiano basata su WORDNET. In *Proceedings of the XXVII Congresso Internazionale di Studi della Società di Linguistica Italiana "Linguaggio e Cognizione"*, Roma, Bulzoni.

Magnini, B., Strapparava C., Ciravegna, F., Pianta, E. (1994). Multilingual Lexical Knowledge Bases: Applied WORDNET Prospects. In Proceedings of the Workshop *The Future of the Dictionary*, Grenoble.

Miller, G.A. (ed.). (1990). WORDNET: An on-line lexical database. *International Journal of Lexicography (special issue)*, 3 (4), pp. 235-312.

Miller, G.A. (1995). A lexical database for English. *Communications of the ACM*, 38(11), pp. 39—41.

Palazzi F. e Folena G. (1992). *Dizionario della lingua italiana*, Torino, Loescher.

Pustejovsky, J. (1995). *The Generative Lexicon*. The MIT Press, Cambridge, Massachusetts.

Siegel S. & Castellan N.J. (1988). *Nonparametric Statistics for the Behavioural Sciences*. McGraw-Hill, second edition.