WordNet for Italian and Its Use for Lexical Discrimination

Alessandro Artale, Bernardo Magnini and Carlo Strapparava
IRST, I-38050 Povo TN, Italy
e-mail: {artale|magnini|strappa}@irst.itc.it

Abstract. We present a prototype of the Italian version of WordNet, a general computational lexical resource. Some relevant extensions are discussed to make it usable for parsing: in particular we add verbal selectional restrictions to make lexical discrimination effective. Italian WordNet has been coupled with a parser and a number of experiments have been performed to individuate the methodology with the best trade-off between disambiguation rate and precision. Results confirm intuitive hypothesis on the role of selectional restrictions and show evidences for a WordNet-like organization of lexical senses.

1 Introduction

WordNet is a thesaurus for the English language based on psycholinguistics principles and developed at the Princeton University by George Miller [Miller, 1990]. It has been conceived as a computational resource, so improving some of the drawbacks of traditional dictionaries, such as the circularity of the definitions and the ambiguity of sense references. Lemmas (about 130, 000 for version 1.5) are organized in synonyms classes (about 100, 000 synsets).

The more evident problem with WordNet is that it is a lexical knowledge base for English, and so it is not usable for other languages. Here we present the efforts made in the development of the Italian version of WordNet [Magnini and Strapparava, 1994; Magnini *et al.*, 1994], a project started at Irst about one year ago in the context of Ilex [Delmonte *et al.*, 1996] a more general project aiming at the realization of a computational dictionary for Italian[1].

A second problem with WordNet is that it needs some important extensions to make it usable for effective parsing. In particular, parsing requires a powerful mechanism for lexical discrimination, in order to select the appropriate lexical readings for each word in the input sentence. In this paper we also explore the integration of "*selectional restrictions*", a traditional technique used for lexical discrimination, with Italian WordNet. Selectional restrictions provide explicit semantic information that the verb supplies about its arguments [Jackendoff, 1990], and should be fully integrated into the verb's argument structure.

Although selectional restrictions are different in different domains [Basili *et al.*, 1996] we are interested in finding common invariants across sub-languages. It

[1] The Ilex consortium includes the Computer Science Department of the University of Torino, the University of Venezia, and the branch of the University of Torino at Vercelli.

is our intention to build a very general instrument that can be afterwards tuned to particular domains by identifying more specific uses. The main motivation is to have both a robust and a computational efficient natural language system. On one hand, robustness is emphasized because sentences that are syntactically correct, but which are not successfully analyzed in the specific application domain, can have a valid linguistic meaning. On the other hand, we are able to filter the sentence meanings on a linguistic basis. This phase discards the unplausible readings pruning the search space by looking for compatibility semantic relations. This kind of discrimination can be realized with computationally effective algorithms by exploiting the lexical taxonomy of WordNet, postponing more complex and expensive computations to the domain specific analysis.

The paper is structured as follow. Section 2 describes the Italian prototype of WordNet; while section 3 shows how selectional restrictions has been added to verb senses. Section 4 shows how Italian WordNet has been coupled with the parser, both for describing lexical senses and as a repository for selectional restrictions. Section 5 reports a number of experiments that has been performed to individuate the methodology design with the best trade-off between disambiguation rate and precision. Finally section 6 provides some conclusive remarks.

2 The Italian WordNet Prototype

The Italian version of WordNet is based on the assumption that a large part of the conceptual relations defined for English (about $72,000$ Isa relations and $5,600$ Part-of relations) can be shared with Italian. WordNet can be described as a lexical matrix with two dimensions: the lexical relations, which hold among words and so are language specific, and the conceptual relations, which hold among senses and that, at least in part, we consider independent from a particular language. The Italian version of WordNet aims at the realization of a multilingual lexical matrix through the addition of a third dimension relative to the language. Figure 1 shows the three dimensions of the matrix: (a) words in a language, indicated by $\mathcal{W}_j$; (b) meanings, indicated by $\mathcal{M}_i$; (c) languages, indicated by $\mathcal{L}_k$. From an abstract point of view, to develop the multilingual matrix it is necessary to re-map the Italian lexical forms with corresponding meanings ($\mathcal{M}_i$), building the set of synsets for Italian (making explicit the values for the intersections $\mathcal{E}^I_{ij}$). The result will be a complete redefinition of the lexical relations, while for the semantic relations, those originally defined for English will be used as much as possible.

An implementation of the Multilingual lexical matrix has been realized which allows a complete integration with the English version and the availability of all the translations for the Italian lemmas. The architecture is easily extendible to other languages. The integration with the computational lexicon ILEX is under development: it will make the access to other levels of lexical information, such as morphological classes, syntactic categories and sub-categorization frames available. The Italian version of WordNet, in December 1996, included about $10,000$ lemmas ($7,000$ nouns, 700 verbs, $1,500$ adjectives, 600 adverbs).

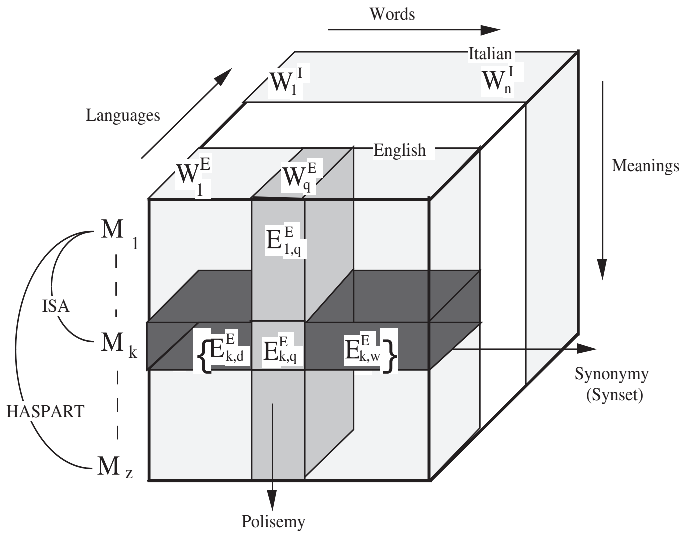

Fig. 1. Multilingual lexical matrix

Till now, data acquisition has been mostly manual, with the help of a graphical interface; however a basic goal of the project is the experimentation of techniques for the (semi)automatic acquisition of data. Algorithms for the resolution of the ambiguities in the coupling with the English WordNet have been developed. Versions automatically created are then tested against manually acquired data, with the aim of incrementally improve the precision level. A final manual check is performed for all the data automatically acquired. It is also foreseen the use of corpora to extract contextual information to be used during the disambiguation process.

3 Adding Selectional Restrictions to Verbs

A number of steps have been followed to add selectional restrictions to Italian WordNet. First, Italian verb senses were extracted from a paper version of an Italian dictionary and checked against a corpus of generic Italian texts. Each verb sense has been then coupled with one or more English WordNet synsets[2]. This phase has been performed manually with the help of a graphical interface (see figure 2) that includes four integrated working tools: (i) a bilingual dictionary with more than 30,000 lemmas; (ii) a graph that allows the visualization of the coupling with the English WordNet; (iii) the bilingual WordNet, that behaves exactly like the English version with the additional possibility to browse the Italian semantic network; (iv) finally, the working cards allow the insertion,

[2] As for figurative uses, they can also be coupled with WordNet provided that an appropriate synset does exist.

Fig. 2. The Italian WordNet interface.

modification and check of the data for a synset. The result of this phase is the extension of the English WordNet with the Italian synsets. Figure 3 shows the correspondence between English and Italian synsets for the verb `Scrivere` (`Write`).

The next step is the definition of the sense subcategorization frame. This includes both syntactic information (i.e., argumental positions, prepositions on indirect objects, category type) and semantic information, such as thematic roles and selectional restrictions. Syntactic information are associated to single verbs, while semantic information are associated to the whole synset, i.e., semantic participants are shared among all the verbs belonging to the synset.

We built selectional restrictions using the synsets of the noun hierarchy. Two different possibilities for defining selectional restrictions are considered:

1. Selectional restrictions obtained from the frames currently provided by WordNet.
2. Selectional restrictions obtained from the whole WordNet noun hierarchy.

As far as the first hypothesis is concerned, WordNet describes all the English verbs resorting to a set of 35 different syntactic frames, which in turn

Synset Label	Italian Synset	English Synset
`Write`	{scrivere redigere comporre}	{write compose pen indite}
`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}

Fig. 3. Correspondences between Italian and English synsets for the verb '*scrivere*' (write).

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 patterns, where the dots can be substituted by the verb stem:

`Somebody ...s`

`Somebody ...s Something`

The problem arising in using these two restrictions is that they are completely uncorrelated to the noun synsets, then, they have to be matched with the proper synsets in the noun hierarchy. The concept `Somebody` includes not only the synset `Person` but also all the synsets denoting group of people that could hold the agent thematic role. We defined `Somebody` using the following boolean combination of synsets:

$$\begin{aligned}\texttt{Somebody} \doteq\ & \texttt{Person} \vee \texttt{People} \vee \texttt{People-Multitude} \vee \\ & (\texttt{Social-Group} \wedge \neg(\texttt{Society} \vee \texttt{Subculture} \vee \\ & \qquad \texttt{Political-System} \vee \texttt{Moiety} \vee \texttt{Clan}))\end{aligned}$$

`Something` is defined as the complement of `Somebody`.

In the second hypothesis selectional restrictions are taken from the whole noun hierarchy. As an example, figure 4 illustrates the senses for the Italian verb `Scrivere` (`Write`) found in Italian WordNet. For each sense we report a conventional name – which unambiguously identifies the synset – and the argumental positions admitted for that sense, with the indication of the selectional restrictions. 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.

A problem is posed by those verb usages in which the argumental position is filled by a synset higher in the hierarchy than the selectional restriction, as in *Cosa hai scritto? (What did you write?)*. For these cases two alternative solutions are possible: (i) extending the simple subsumption check to a more comprehensive "double subsumption" check, in which or the filler is subsumed by the restriction, or the restriction is subsumed by the filler. This approach has been successfully experimented in a number of prototype systems developed at Irst. (ii) considering such high level fillers as pronouns, which stand for a noun satisfying the verb selectional restrictions.

Finding the appropriate selectional restrictions revealed itself difficult and time consuming. The process required a deep search into the WordNet noun hierarchy. 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 experimental results. This iterative process ended with complex selectional restrictions for verbs, as the figure 4 shows.

The WordNet verb taxonomy is based on the *troponymy* relation, which is defined as the co-occurrence of both lexical implication and temporal co-extension between two verbs. We would note that, every time a troponymy relation between two verbs holds, an Isa relation between the correspondent selectional restrictions holds, too.

WordNet **Synset**	**Subject**	**Object**	**Indirect-Object**
`Write`	`Somebody`	`Written-Material` ∨ `Symbolic-Repres` ∨ `Saying` ∨ `Correspondence` ∨ `Sentence` ∨ `Message` ∨ `Message-Content` ∨ `Code` ∨ `Symbol` ∨ `Date` ∨ `Language-Unit` ∨ `Property` ∨ `Address-Speech` ∨ `Print-Media`	−−
`Write-Music`	`Person`	`Music`	−−
`Write-Communicate`	`Somebody`	(`Written-Material` ∧ ¬`Section`) ∨ `Symbolic-Repres` ∨ `Saying` ∨ `Sentence` ∨ `Name` ∨ `Message` ∨ `Message-Content` ∨ `Code` ∨ `Date` ∨ `Property`	`Somebody`
`Write-Publish`	`Somebody`	`Written-Material` ∨ (`Print-Media` ∧ ¬`Section`)	`Print-Media` ∨ `Publishing-House`
`Write-Send`	`Somebody`	`Correspondence` ∨ `Message` ∨ `Letter-Missive`	`Somebody`

Fig. 4. Lexical entries for `Scrivere` (`Write`).

4 Coupling WordNet and a TFS Parser

In this section we describe the architecture we used for checking WordNet usability in parsing. Italian WordNet has been used in two different phases of the linguistic analysis. On a first phase, we use Italian WordNet as a lexicon repository to carry on lexical analysis. During the semantic analysis Italian WordNet is used as a kind of Knowledge Base (KB) exploiting the structural relationships among synsets. In particular, we used the supertype/subtype-like hierarchy of synsets during the parsing process in order to discard unplausible constituents on a semantic base.

The parser used is a CYK chart parser embedded in the Geppetto environment [Ciravegna *et al.*, 1996], and coupled with a proper unification algorithm. Geppetto is based on a Typed Feature Logic [Carpenter, 1992] for the specification of linguistic data. The Geppetto environment allows to edit and debug grammars and lexica, linking linguistic data to a parser and/or a generator, integrating various form of KBs, and using specialized processors (e.g., morphological analyzers). In particular, we integrated the hierarchical structure of WordNet as an external KB, while an Isa function uses the WordNet hierarchy in order to check subsumption relationships between WordNet synsets.

The grammar is written adopting a HPSG-like style, and each rule is regarded as Typed Feature Structure (TFS). For the current experiment the grammar coverage is limited to very simple verbal sentences formed by a subject, a main verb together with its internal arguments and, possibly, an adjunct phrase. Observe that, the syntactic analysis does not take into account the pp-attachment case. We excluded the possibility to capture these complex nominal phrases. Indeed, the object of the experiment is to disambiguate among WordNet senses of both verbs and nouns on the basis of the lexical semantic restrictions for the arguments of the verb and the lexical semantic associated to the noun.

A condition for using WordNet coupled with the Geppetto environment is to bring it in a format effectively usable. The exploited idea was to rebuild the WordNet hierarchy in Clos, the object-oriented part of Common Lisp. The advantages of this approach is the possibility to implement a fast and flexible access to the synsets hierarchy and, in particular, an efficient Isa functionality as required for the semantic checking during the parsing. The arguments to Isa function may be a complex boolean combination of synsets (e.g., see selectional restrictions in figure 4).

The parser controls the overall processing. Whenever it tries to build a (partially recognized) constituent it incrementally verifies the admissibility of the semantic part of such a constituent, using the WordNet hierarchy. In particular, whenever a noun is associated with a verbal argument the Isa function is triggered to check whether the synset of the noun is subsumed by the selectional restriction of the corresponding verbal argument. Due to the large number of analyses, it is useful to discard unplausible constituents as soon as possible to cut the search space. This has been obtained interliving the syntactic and semantic processes: as soon as the semantic test fails the constituent is rejected.

Word	WordNet SynsetLabels
`Regina (Queen)`	`Queen-Insect,Queen-Regnant,Queen-Wife,` `Queen-Card,Queen-Chess`
`Articolo (Article)`	`Article-Artifact,Article-Clause,` `Article-Grammar,Article-Document`
`Lettera (Letter)`	`Letter-Missive,Letter-Alphabet`
`Libro (Book)`	`Book-Publication,Book-Section,Book-Object`

Fig. 5. Lexical entries for nouns.

5 Experiments and Results

In this section we describe the empirical results obtained by coupling a WordNet based lexicon with a parser. In our intention, the experiment should bring evidences for the following aspects:

- Plausibility of WordNet senses for describing lexical entries;
- Usability of WordNet for carrying out lexical discrimination.

The experiment has been carried out on 60 sentences with 1201 different readings, and formed by using seven verbs (*write, eat, smell, corrode, buy, receive, associate*) coupled with fifty common nouns and two proper nouns. In the general experimental setting a sentence is given to the parser in a situation characterized by multiple lexical entries for each single word (one for each WordNet sense). The analyses produced by the parser are compared with the set of interpretations given by a human.

As far as nouns are concerned, a lexical entry includes all the senses found in Italian WordNet. Some of the nouns used in the experiment are shown in figure 5. As for verbs, we started from the Italian WordNet senses and then we faced to the problem of individuating the proper selectional restrictions for each argumental position of the verb subcategorization frame as seen before. So we build a small number of lexical entries, by means of which we composed the sentences of the experiment. We experimented the two hypotheses on selectional restrictions presented in section 3, i.e., the one with general WordNet frames and the other with more refined selectional restrictions.

As an example, figure 6 shows the output of the parser for the sentence "*La regina scrisse una lettera a Giovanni*" ("*The queen wrote a letter to John*"). As a convention, internal arguments are represented by the symbol '/', while a '//' denotes a verbal adjunct. This sentence was selected because it produces an high number of readings (40) among the test suite sentences. This is due to both the verb sense ambiguity (`write` has five senses) and to the noun ambiguities (`queen` has five senses, and `letter` two). Note that the parser excludes the

Sentence	`La regina scrisse una lettera a Giovanni`	
	(THE QUEEN WROTE A LETTER TO JOHN)	
Restrictions	No semantic discrimination	
Number of readings	40	
Restrictions	Discrimination with WORDNET Frames (I exp. setting)	
Number of readings	16	
Restrictions	Discrimination with WORDNET Full Hierarchy (II exp. setting)	
Number of readings	8	
Readings	$(1, 2, 3, 4, 5, 6)$	
	Queen-Wife/Write/Letter-Alphabet//John	(7)
	Queen-Regnant/Write/Letter-Alphabet//John	(8)
Human Judgment		
Number of readings	6	
Readings	Queen-Regnant/Write-Communicate/Letter-Missive/John	(1)
	Queen-Wife/Write-Communicate/Letter-Missive/John	(2)
	Queen-Regnant/Write-Send/Letter-Missive/John	(3)
	Queen-Wife/Write-Send/Letter-Missive/John	(4)
	Queen-Regnant/Write/Letter-Missive//John	(5)
	Queen-Wife/Write/Letter-Missive//John	(6)

Fig. 6. An example of sentence

sense `Write-Publish` since the indirect object must be introduced by the Italian prepositions "*su*" or "*per*" (in English "*on*" or "*for*"), while in this example we have the preposition "*a*" ("*to*").

Let us first consider the results obtained in the second experimental setting, which best approximates the human judgment. Out of the eight interpretations accepted, two are implausible for a human reader. This is caused by the contemporary presence of the sense `Letter-Alphabet` and of the proper noun `John` as, respectively, patient and beneficiary of the `Write` verb sense. Note that, each of these senses are, per se, valid arguments since they satisfy the selectional restrictions.

In the first experimental setting, the presence of weaker selectional restrictions (just *Somebody*, *Something*) yields more spurious readings. As a matter of fact, the more evident problem is that in many cases argumental positions are not properly filled. For example, a reading is allowed in which "a `Queen-Regnant` can `Write-Music` a `Letter-Missive`" (i.e., a kind of correspondence).

Figure 7 reports the quantitative results of the experiment. They are preliminary since they have been obtained on a limited number of sentences (60). For each experimental setting the number of total readings produced by the parser, the discrimination rate (i.e., the rate of the rejected readings: $(1201 - x)/1201$), and the precision (i.e., the rate of correct readings: $122/x$) are shown. 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` an `Article-Clause` on a `Newspaper-Periodic`); (ii) aspectual properties of the verb (e.g., it is very difficult to interpret *La regina sta scrivendo un articolo sul gionale* (*The queen is writing an article on the newspaper*) with the `Write-Publish` sense, because publishing is a culminative process).

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

Fig. 7. Quantitative results obtained on 60 sentences

6 Conclusions

In this paper we presented the approach underlying the Italian WordNet, a general computational lexical resource. A prototype has been realized which implements a multilingual lexical matrix. In light of the concrete use of Italian WordNet we propose the integration of selectional restrictions into the verbal taxonomy. The acquisition of selectional restrictions for the present experiment has been manual with the help of a graphical interface. For the future it will be necessary to consider the possibility of automatically extract selectional restrictions from corpora by means of already known techniques (e.g. [Basili *et al.*, 1996]).

The empirical verification which has been performed confirms the intuitive hypothesis that selectional restrictions crucially affect lexical disambiguation and that the discrimination rate improves as far as they are more detailed. Some general suggestions can be drawn in order to individuate a trade-of between the effort necessary for describing selectional restrictions and the lexical disambiguation obtained. Although the definition of detailed selectional restrictions was highly time consuming, our experience shows that this approach obtains good results both in the discrimination rate and in the precision.

The experiment also brings evidence for a WordNet like sense organization. In fact, different selectional restrictions apply to different senses allowing to discriminate among different readings. However, an important drawback in WordNet is the lack of relations among related senses of the same word. This is particularly crucial for the *logical polysemy* cases [Pustejovsky, 1995], when a sense can be generated from another in a predictable way, and, in general, to treat the so called "verb mutability effect" as discussed in [Gentner and France, 1988].

References

[Basili *et al.*, 1996] R. Basili, M.T. Pazienza, and P. Velardi. Integrating general-purpose and corpus-based verb classification. *Computational Linguistics*, 22(4), 1996.

[Carpenter, 1992] B. Carpenter. *The logic of typed feature Structures.* Cambridge University Press, Cambridge, Massachusetts, 1992.

[Ciravegna *et al.*, 1996] F. Ciravegna, A. Lavelli, D. Petrelli, and F. Pianesi. The Geppetto environment, Version 2.0.b. User Manual. Technical report, IRST, 1996.

[Delmonte *et al.*, 1996] R. Delmonte, G. Ferrari, A. Goy, L. Lesmo, B. Magnini, E. Pianta, O. Stock, and C. Strapparava. ILEX - un dizionario computazionale dell'italiano. In *Proc. of 5 th Convegno Nazionale della Associazione Italiana per l'Intelligenza Artificiale*, Napoli, September 1996.

[Gentner and France, 1988] D. Gentner and I.M. France. The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs. In S. Small, G.W. Cottrell, and M.K. Tanenhaus, editors, *Lexical Ambiguity Resolution.* Morgan Kaufman, San Mateo, California, 1988.

[Jackendoff, 1990] Ray Jackendoff. *Semantic Structures.* Current Studies in Linguistics. The MIT Press, Cambridge, Massachusetts/London, England, 1990.

[Magnini and Strapparava, 1994] B. Magnini and C. Strapparava. Costruzione di una base di conoscenza lessicale per l'italiano basata su WordNet. In *Proc. of the 28 th International Congress of the Società Linguistica Italiana*, Palermo, Italy, Ottobre 1994.

[Magnini *et al.*, 1994] B. Magnini, C. Strapparava, F. Ciravegna, and E. Pianta. Multilingual lexical knowledge bases: Applied WordNet prospects. In *The Future of Dictionary - Workshop sponsored by Rank Xerox European Research Centre and ESPRIT Project Acquilex II*, Grenoble, France, October 1994.

[Miller, 1990] G. A. Miller. WordNet: "An on-line lexical database". *International Journal of Lexicography (special issue)*, 3(4):235–312, 1990.

[Pustejovsky, 1995] J. Pustejovsky. *The Generative Lexicon.* The MIT Press, Cambridge, Massachusetts, 1995.