SENSEVAL-3: Third International Workshop on the Evaluation of Systems
for the Semantic Analysis of Text, Barcelona, Spain, July 2004
Association for Computational Linguistics

Senseval-3: The Italian All-words Task

Marisa ULIVIERI
ILC-CNR
Via Moruzzi 1
Pisa, Italy, 56124
marisa.ulivieri@ilc.cnr.it

Elisabetta GUAZZINI
ILC-CNR
Via Moruzzi 1
Pisa, Italy, 56124
elisabetta.guazzini@ilc.cnr.it

Francesca BERTAGNA
ILC-CNR
Via Moruzzi 1
Pisa, Italy, 56124
francesca.bertagna@ilc.cnr.it

Nicoletta CALZOLARI
ILC-CNR
Via Moruzzi 1
Pisa, Italy, 56124
glottolo@ilc.cnr.it

Abstract

This paper describes the Italian all-words sense disambiguation task for Senseval-3. The annotation procedure and criteria together with the encoding of multiwords are presented.

1 Introduction

This paper describes the Italian all-words sense disambiguation task for Senseval-3: about 5000 words were manually disambiguated according to the ItalWordNet (IWN) word senses. The first section briefly describes of the corpus and the lexical reference resource. The second section contains some general criteria adopted for the annotation of the corpus and illustrated by a series of examples. Issues connected to the treatment of phenomena typically found in corpora, e.g. abbreviations, foreign words, jargon, locutions are discussed. Furthermore, the encoding of compounds, metaphorical usages, and multiword units is described. Problems connected with i) the high granularity of sense distinctions in the lexical resource and ii) unsolvable ambiguities of the contexts are dealt with. Finally, it is evidenced how the annotation exercise can be of help in updating or tuning IWN, by adding missing senses and/or entries.

2 The Corpus and the Lexical Resource

The Italian all-words corpus consists of about 13600 word tokens, extracted from the SI-TAL [1], Italian Syntactic Semantic Treebank (ISST). The ISST (Montemagni *et al*. 2003) consists of i) a generic corpus of about 215,000 tokens, extracted from different periodicals and newspaper articles (*La Repubblica*, *Il Corriere della Sera*) and ii) a specialised corpus of about 90,000 tokens, with texts belonging to the financial domain (*Il Sole-24Ore*). The annotated corpus consists of about 5000 words and comprises a selection of Italian newspaper articles about various topics: politics, sport, news, etc. The common data format is XML.

The reference lexical resource used for the Senseval-3 sense tagging task is the lexical-semantic database IWN, developed within the framework of two different research projects: EuroWordNet (Vossen 1999) and SI-TAL, during which IWN was extended by the insertion of adjectives, adverbs and a subset of proper nouns. The IWN database contains about 64,000 word senses corresponding to about 50,000 synsets. It has inherited the EWN linguistic model (Alonge e*t al*., 1998) which provides a rich set of semantic relations, and the first nucleus of verbs and nouns. IWN was structured around the notion of *synset*, a set of synonymous word meanings, and the information is encoded in the form of lexical-semantic relations between pairs of synsets. The IWN database comprises also an Interlingual Index (ILI), based on the Princeton WordNet 1.5 used to link wordnets of different languages so that it is possible to go from the Italian words to similar words in English language. IWN has also inherited from EWN the Top Ontology (TO), a hierarchical structure of language-independent concepts, reflecting fundamental semantic distinctions. Via the ILI, all the concepts in the wordnet are linked to the Top Ontology.

[1] SI-TAL (Integrated System for the Automatic treatment of Language) was a National Project devoted to the creation of large linguistic resources and software tools for Italian written and spoken language processing.

3 Annotation Procedure and Criteria

For the Italian all-words task , the annotation was carried out manually, word by word following the text. For each word, annotators were supplied with information about The tagging operation consisted in the assignment of a sense number to each full word or sequence of words corresponding to a single unit of sense, such as compounds, idioms, metaphorical usages, etc. The sense number which refers to a specific synset was assigned by the annotators according to the lexical resource IWN. The assignment of a sense number allows tagged words to inherit a series of semantic information ranging from meronymy, synonymy, hyperonymy, etc. up to the fundamental semantic distinctions of the Top Ontology.

The annotation of the corpus was restricted to nouns (2583), verbs (1858), adjectives (748), a group of multiword expressions (97 – verb phrases, adjectival phrases and noun phrases) and a set of general proper nouns (163). Two linguists disambiguated the texts. The annotators made every effort to match a text word to a IWN sense, but sometimes this could not be done, since the required sense was not present in the reference resource. Cases of difficult sense attribution and of disagreement between annotators were marked and left to further discussion and refinement. Frequently, a tight interaction between the IWN developers and the annotators was needed. By the way, this collaboration produced a double-sided effect: on the one side, the lexical resource gained in coverage, being enlarged through the addition of missing entries and/or senses and, on the other side, the corpus encoding has been made possible.

3.1 Non-annotated cases

Notwithstanding this, some cases have been left "empty". They are in particular terms with not standard meaning, often absent from dictionaries[2] as, for example: i) abbreviations (*C.T. Commissario Tecnico,* Technical Officer); ii) foreign words (e.g. *off limits*); iii) jargon (e.g. *fumantino* adj. *una persona fumantina*, an irascible person); iv) terms semantically modified through evaluative suffixation (e.g. *vetturetta*, small car); v) locutions (e.g. *per carità!*, for goodness' sake!*; ci mancherebbe*, that's all we need); vi) words, or sequences of words, indicating human association groups (*Caschi blu*, the Blue Berets, *Croce Rossa*, Red Cross, etc.), vii) nicknames (*Scarpa d'oro*, lit. Gold Shoe – to say a good football player, *Primula Rossa*, the Scarlet Pimpernel, a mafioso boss, etc.); viii) neologisms (e.g. *komeinista berlusconiano,* concerning Khomeini, Berlusconi).

This type of specific neologisms or idiomatic expressions have a high frequency in corpora. Corpus texts are extracted from newspaper articles about politics, sports, news, etc. in which a high number of words currently used in the everyday language of media appear. Rarely a lexical resource contains this new-born expressions hence not completely meeting the requirements of semantic annotation.

3.2 Fully-compositional Expressions

It may be the case that annotators had to deal with complex expressions where the meaning is compositional, e.g. *Ministero della Difesa* (Department of Defence). Even if this sequence of words could be perceived by native speakers as a single multiword expression, the reference lexical resource did not present it as an independent entry. This is a case of fully compositional expression, whose interpretation depends functionally on the interpretation of its constituents. They were, therefore, decomposed and annotated according to their parts.

3.3 Metaphorical Usages

Figurative and metaphorical usages were hard to map to the correct sense: sometimes, it has been necessary to accept for them, at least, a compromise solution. Consider the following context (where bold marks the figurative usage):

due lavoratori su tre **sono a casa** = essere disoccupato

out of three workers, two are at home = to be unemployed)

The interpretation of the context presupposes an extra-linguistic knowledge, which cannot be encoded at the lexical-semantic level of description, even if the collocation with *lavoratori* (workers) allows to correctly disambiguate. In this case *a* (at home) represents an instance of a non lexicalised metaphor, therefore it was not possible to assign the appropriate figurative sense. A compromise solution was adopted and the individual components of the phrase were annotated, even if the correct semantic contribution of the multiword expression was lost.

Another interesting case is provided by the occurrences of some metaphoric uses of verbs. Consider these examples:

[2] Some of them are very technical-specialistic terms or expressions extracted, in particular, from the soccer domain, e.g. *andare in percussione/in sovrapposizione* (*lit*. to go in percussion/in overlapping).

… è **andata male** in Spagna e non si è qualificata alle …, *lit.* in Spain it went badly and did not qualify for the

… il rapporto **andò avanti** fino alle nozze, *lit*. the relationship went ahead until wedding

In the first example, even if the verb is frequently used with this meaning, it was not possible to attribute a correct sense number, since it was not accounted for in the lexical resource.

In the second case, the verbal locution *andare avanti* was not present in IWN. In this context, *andare* has been annotated with $andare_{11,}$ 'to progress', which incorporates the meaning provided by *andare* plus the adverb *avanti* (to go ahead).

All the above mentioned cases of non-annotation or compromise annotation evidence the divergences between lexicon encoding, on the one hand, in which senses are by necessity "decontextualised" to be able to capture generalizations (Calzolari *et al*. 2002) and corpus annotation, on the other, where "contextualization" plays a predominant role and, consequently, figurative senses, idioms, metaphorical usages, multiwords, are highly frequent.

3.4 High granularity of sense distinctions

One of the main reasons for disagreement between annotators could arise from the high IWN granularity in sense distinction. Often, when deciding a sense, too subtle distinctions could turn out to be a disadvantage for the annotators. Consider the verb *sentire* (to hear): IWN makes a very fine-grained distinction, where exactly 17 senses are available. Some of them overlap or are so close each other to be undistinguishable for human annotators and may be problematic for systems. In the following example:

Passano pochissimi secondi e qualcuno sente un urlo= … and someone hears a cry

In IWN, two senses of the verb are overlapping:

sentire 1 – percepire con l'orecchio (to hear)

sentire 2 – essere in grado di percepire suoni con l'orecchio (to be able to hear sounds)

These distinctions are too subtle to be used in corpus annotation. The annotator has chosen sense1, but (in order to allow for the coarse-grained scoring of automatic annotation) a *sensemap* of words, a table where the overlapping senses are accounted for, was provided.

3.5 Context Ambiguity

Corpus annotation strategy allowed to handle cases where synsets are numerous and present fine-grained distinctions, not easily mappable to the corpus contexts or cases in which the context could raise a double interpretation. Annotators were not forced to make a totally subjective choice and could assign multiple senses ('and' operator).

Lo Zaire è uno dei **paesi** più pericolosi di tutta l'Africa = The Zaire is one of the most dangerous countries of Africa

In IWN, sense distinctions are as follows:

Paese1– territorio con un governo sovrano e una propria organizzazione politica e amministrativa, (territory with its own political and administrative organization)

Paese3– insieme di individui legati da stesse tradizioni storiche, lingua, costumi, (group of people with same historical traditions, languages and customs)

Since annotators could not achieve a satisfactory disambiguation, they take into account both senses, sense1 and sense3. It was not clear, indeed, if the dangerousness refers to the country (sense1) or to the people (sense3). During the annotation, multiple senses have been assigned to about 90 lemmas, that appeared arbitrary or impossible to disambiguate.

3.6 Multiwords annotation

The main difference between Senseval-3 and Senseval-2 is that in the all-words annotation task annotators are faced with complex lexical items coined, generally, with many technical terms, collocations, idioms, compounds, frozen expressions or multiwords, which were not present in the lexical-sample task. With the term *multiwords* we refer to all sequences of words whose meaning cannot be built compositionally from the meanings of its component words. The semantic contribution of individual components if annotated separately does not give reason of the final meaning of the expression which can be considered a sort of "new concept", e.g. *farla franca* (to get away with it), *prendere parte* (to take part), *muro a secco* (dry-stone wall).

In IWN a set of lexicalised expressions were already included and the correct sense to assign was, hence, available: *perdere i sensi* (to faint), *fare fuoco* (to fire), *passare in rivista* (to review) etc. Many multiword expressions found in the corpus were added to IWN as semantically complex units, e.g. *vedersela brutta* (to have a

narrow escape), *essere in corso* (to be in progress) etc. The annotation task has given, hence, the opportunity to establish a strong interaction between annotators and lexicographers in deciding what kind of sequences were real multiword expressions and, above all, which were worthwhile from a linguistic point of view to introduce in the lexical resource.

The multiword expressions (about 60) were annotated with the following information: multiword ID; Part_of_Speech lemma; function of the components words: head, satellite.

The individuation of the headword of the sequence has been made on the basis of a lexical criterion: for noun-phrases, the head of the sequence was considered the noun, the adjective for adjectival–phrases, the verb for verb-phrases. Once recognized the head, the other constituents play the role of satellites and the whole sequence receives the part-of-speech of the head.

Here is an example encoded in XML:

```
<head id="cs.morph074.mw_704"
sats="cs.morph074.mw_706
cs.morph074.mw_707">
uscita</head>clamorosamente
<sat id="cs.morph074.mw_706">di</sat>
<sat id="cs.morph074.mw_707">scena</sat>
<answer head="cs.morph074.mw_704"
senseid="uscire di scena.V.1"/>
```

Our intention was only to provide an examplification of the methodology we adopted when trying to handle multiword expressions. Even if the recognition and treatment of poly-lexical units is obviously one of the most important issues emerging in the process of context interpretation, in this paper we did not address theoretical issues concerning their exact identification.

References

Alonge Antonietta, Nicoletta Calzolari, Piek Vossen, Laura Bloksma, Irene Castellon, Maria Antonia Marti and Wim Peters. 1998. The Linguistic Design of the EuroWordNet Database. *Special Issue on EuroWordNet. Computers and the Humanities*, Vol.32, N. 2-3: pag.91-115.

Calzolari Nicoletta, Claudia Soria, Francesca Bertagna and Francesco Barsotti. 2002. Evaluating Lexical Resources Using Senseval. Journal of Natural Language Engineering, Special Issue of Senseval-2, vol.VIII(4), pag. 375-390.

Kilgarriff Adam, Rosenzweig Joseph. 2000. *English Senseval: Report and Results*. Proc. Second Conf on Language Resources and Evaluation *Athens,* pag. 1239-1244.

Montemagni Simonetta, Barsotti Francesco, Battista Marco Calzolari Nicoletta, Corazzari Ornella, Lenci Alessandro, Pirrelli Vito, Zampolli Antonio, Fanciulli Francesca; Massetani Maria, Raffaelli Remo, Basili Roberto, Pazienza Maria Teresa, Saracino Dario, Zanzotto Fabio, Mana Nadia, Pianesi Fabio, Delmonte Rodolfo. 2003. The syntactic-semantic *Treebank* of Italian. An Overview. *Linguistica Computazionale a Pisa vol*. I, pag.461-492.

Vossen Piek (ed.) 1999. *EuroWordNet General Document*. The EWN CD-ROM (see also: http://www.hum.uva.nl/).