The Italian Lexical Sample Task

Francesca BERTAGNA
Consorzio Pisa Ricerche
Via S. Maria 40
56100 Pisa, Italy,
f.bertagna@ilc.pi.cnr.it

Claudia SORIA
Istituto di Linguistica Computazionale-CNR
Via Moruzzi 1
56100 Pisa, Italy,
soria@ilc.pi.cnr.it

Nicoletta CALZOLARI
Istituto di Linguistica Computazionale-CNR
Via Moruzzi 1
56100 Pisa, Italy,
glottolo@ilc.pi.cnr.it

Abstract

In this paper we give an overall description of the Italian lexical sample task for SENSEVAL-2, together with some general reflections about on the one hand the overall task of lexical-semantic annotation and on the other about the adequacy of existing lexical-semantic reference resources.

Introduction

In this paper we give an overall description of the Italian lexical sample task for SENSEVAL-2. In the first two sections, the corpus and reference lexicon used are illustrated; the last section contains some general reflections on the basis of the Senseval experience about on the one hand, the overall task of lexical-semantic annotation and on the other, about the adequacy of existing lexical-semantic reference resources.

Dictionary and Corpus

The dictionary and corpus used for the Italian lexical sample task were provided by the resources developed in the framework of the SI-TAL project[1]. The data had not been adapted in order to be used for the Senseval task, apart from the necessary format conversions. A common encoding format (XML) proved to facilitate re-use and sharing of the data.

The lexical sample corpus

The Italian lexical sample corpus (test data only) consisted of about 3900 instances for 83 lexical entries (46 nouns, 21 verbs, and 16 adjectives), with an average of 47 contexts per entry.
The lexical samples were taken from the SI-TAL Italian Syntactic-Semantic Treebank (ISST[2]), which was still under development when the Senseval task was organized. This fact implied as a main disadvantage of the ISST material that corpus instances were associated with very little context. For each instance, the context corresponded to the sentence containing the target word and in our experience sometimes this proved to be not enough for a WSD task.
The ISST consists of two sub-components: a generic and a domain-specific (financial) corpus, of about 215,000 and 90,000 tokens, respectively. The annotated material comprises instances of newspaper articles, representing everyday journalistic Italian language. As far as annotation is concerned, the ISST has a three-level structure: two levels of syntactic annotation (a constituency-based and a functional-based annotation level) and a lexical-semantic level of annotation. ISST is supposed to be used in different types of applications, ranging from training of grammars and sense disambiguation systems, to the evaluation of language technology systems.
For its use in the SENSEVAL-2 task, only the semantic annotation was used, even if it is

[1] SI-TAL ('Integrated System for the Automatic Treatment of Language') is a National Project, coordinated by Antonio Zampolli at the 'Consorzio Pisa Ricerche' and involving several research centers in Italy, aiming at developing large linguistic resources and software tools for the Italian written and spoken language processing.

[2] See Montemagni et al. (2000a) and Montemagni et al. (2000b).

conceivable that a system could make use of the syntactic information as well.

In the ISST, this was performed manually using the ItalWordNet lexicon (henceforth IWN, see Roventini et al. 2000) as a reference resource (see below for a description). Semantic annotation consisted in assigning to each full word or sequence of words corresponding to a single unit of sense (such as compounds, idioms, etc.) a given sense number (referring to a specific synset) taken from IWN, plus specific features created for the annotation task to account for idioms, compounds and multi-words, figurative uses, evaluative suffixation, foreign words, proper nouns and titles, among the others. From this point of view, the semantic annotation of the corpus enriches the information available in the lexical resource.

However, in order to comply with the SENSEVAL-2 lexical sample format, the only semantic information used was the sense number of ISST, corresponding to the sense number of IWN synset variants, while the supplementary features had to be discarded. This fact obviously resulted in a loss of the overall semantic information available.

For instance, the semantic annotation gave no information about the specific domain or about possible metaphoric senses.

Although the original ISST contained multiwords expressions, no one of them was included in the Senseval lexical sample.

The selection of the lemmas has been carried out starting from the analysis of part of the words chosen for the English lexical sample, since we wanted to share a minimal overlapping core with the English list, in order to make the final results more comparable in a multilingual perspective". At the end, the overlap between English and Italian consisted of only 8 entries, unfortunately.[3]

The criteria for the selection were the polysemy of the word in the lexicon, the frequency, and the actual occurrence in the annotated resource with more than one meaning.

The average polysemy was of 5 senses per word (5 for the nouns subset, 6 for the verbs and 3 for the adjectives).

The average frequency turned out to be rather low, since the Italian treebank from which the lexical sample was extracted was still not complete and we had to select the most frequent words with at least two senses in the lexicon and used at least in two of their senses in the annotated corpus. This led to select mainly words with quite high polysemy and rather generic senses. For instance, only 12 of the 46 nouns had also concrete sense.

More importantly, since we had at our disposal a rather low number of occurrences, no training data were available for the Italian task. This makes the results for the Italian task hardly comparable with those which used similarly structured data, such as the Spanish, Swedish, Basque and Korean tasks as all of them had training data available. This is particularly significant in evaluating the results for the Italian task if we consider that the two systems participating to the task were supervised and needed sense-tagged training instances of each word. For the next Senseval, a larger annotated corpus will be available and hence a training corpus will be provided.

The reference lexicon.

As it was said before, the occurrences provided for the WSD lexical sample task were annotated according to the lexical-semantic database ItalWordNet, developed within the framework of the SI-TAL Project[4].

ItalWordNet is an extension of the Italian wordnet built during the EuroWordNet project (Vossen, 1999).

The IWN database is constituted by:

i) a generic wordnet containing about 64,000 word senses corresponding to about 49,000 synsets;
ii) a (generic) Interlingual-Index (ILI) which is an unstructured version of WordNet 1.5, also used in EWN to link wordnets of different languages;
iii) a terminological wordnet, containing about 5,000 synsets of the economic-financial domain;
iv) a terminological ILI, to which the terminological wordnet is linked;
v) the Top Ontology, a hierarchy of language-independent concepts, built within EWN and partially modified in IWN to account for adjectives (Alonge et al., 2000). Via the ILIs, all the concepts in the generic and specific wordnets are directly or indirectly linked to the Top Ontology;
vi) the Domain Ontology, containing a set of domain labels. Via the ILIs, all the concepts in the generic and specific wordnets are

[3] The entries that are in common were: *arte-art, chiamare-call, colpire-hit, giocare/gioco-play, lavorare/ lavoro-work, senso-sense, trovare-find.*

[4] ItalWordNet is a joint effort between the Consorzio Pisa Ricerche and IRST (Istituto per la Ricerca Scientifica e Tecnologica), Trento, Italy.

directly or indirectly linked to the Domain Ontology.

For the 83 lexical entries we provided to the competitors a hierarchical basic data structure: all the senses of the lemma organized in groups of synonyms (synset) plus their direct hyperonyms and a brief Italian definition.
We also provided a set of semantic relations (belonging to the set of Euro(/Ital)WordNet relations: hyponymy, role/involved, holo/meronymy, derivational relations etc.), but we didn't supply the target entries of the relations (and all their semantic and ontological information) since we provided only a portion of the whole wordnet[5].
All the entries were provided with equivalence relations to at least one record of the EuroWordNet Interlingual Index and with the link to the EuroWordNet Top Concepts.
The entries have been used as they were in the wordnet, without making any adjustment specific for the task at hand. Although the domain information, so useful in a WSD task, is available in the model (only with few labels), none of the provided entries had it, because it has not been systematically codified and also because almost all the entries were quite generic. This was a main disadvantage for at least one of the two systems competing for the Italian Senseval task.
We are now in the process of evaluating whether a linking between ItalWordNet and SIMPLE[6] would be feasible; such a linking could allow ItalWordNet to inherit the rich domain information available in the SIMPLE database.
We didn't consider the POS-tagging a part of the task and we provided as corpus instances only those with the same POS as the previously selected lexical items, i.e. we eliminated occurrences of homographs belonging to different parts of speech.

Results for the Italian lexical sample task

Only two systems took part in the Italian task, namely the IRST and JHU systems.
The results for fine, mixed and coarse-grained WSD are illustrated in the following tables:

System	Precision	Recall	Attempted
IRST	0.406	0.389	95.783%
JHU	0.353	0.353	100%

Table 1: Fine-grained scoring

[5] The whole of the new version of IWN could be obtained through ELRA.

[6] See Lenci et al. (2000)

System	Precision	Recall	Attempted
IRST	0.482	0.461	95.783%
JHU	0.421	0.421	100%

Table 2: Mixed-grained scoring

System	Precision	Recall	Attempted
IRST	0.483	0.463	95.783%
JHU	0.423	0.423	100%

Table 3: Coarse-grained scoring

The low scores are mainly due to the lack of training data and of domain information. It is also possible that for some entries of the lexicon the subtlety of sense distinctions contributed to low performance of the systems, as it's shown by better results obtained with the coarse-grained scoring.

General remarks

Starting from the SENSEVAL-2 experience, we would like to make a few general remarks, both about the adequacy of available lexical-semantic reference resources for WSD tasks and about the overall task of lexical-semantic annotation.
One of the well-known problems of WordNet is the fine-grainedness of its entries in terms of sense distinction. This is true also for the Italian net, even if maybe at a lower level: a brief analysis of the entries highlights the presence of some very subtle distinctions among the senses. Actually, during the SI-TAL project, corpus annotators set up a specific annotation strategy for handling cases where synsets are numerous and reflect fine-grained sense distinctions not easily mappable to the corpus contexts. The strategy allowed the assignment of multiple senses connected through logical operators of conjunction (when IWN senses cannot be distinguished) vs. disjunction (when the ambiguous context does not allow a choice among the different IWN senses).
Nonetheless, in the Italian lexical sample used for Senseval, there are about only 140 cases of multiple key assignment out of about 3900 corpus instances.
This suggests that vague or too fine-grained distinctions are still unproblematic for humans, but may become problematic for machines. It could be useful to investigate what kind of sense distinctions are hardest for systems to make, and whether or not systems have problems with the same senses that human annotators have problems with.
When a stable version of the annotated resource is available, we will be able to start a more detailed analysis of the results of the annotation.

It will be possible, for example, to evaluate the impact of the presence of figurative/rhetorical nuances of a sense in the corpus or to consider the quality and types of the multiwords that, found in the corpus, have been proposed to the IWN lexicographers in order to have them added to the lexicon.

But, above all, by analysing the level of confidence in the sense assignment, it will be possible to evaluate the correctness/suitability of the sense distinction in those cases that generated doubts in the human annotators. This kind of analysis would be particularly useful under the perspective of the organization of future Senseval tasks.

Another issue to inquiry is whether the adoption of the wordnet model and the use of the synsets as information core can lead to a proliferation of word meanings according to the kind of synonyms which may replace a given word in a context[7]. Apart from this, however, it is a fact that use of wordnet or wordnet-like resources significantly correlates with an overall worsening in the performance of WSD systems compared with the previous results obtained using traditional dictionaries. This certainly is an issue to reflect upon.

Other, more general considerations concern the issue of semantic annotation in general. It does not seem correct to talk about the "right sense distinction", and to think at the word sense as a task–independent information (Kilgarriff, 1997): the greater vs. lesser granularity depends also on the task/domain/situation and in principle there is no upper or lower limit to sense granularity.

It seems that there are areas of meaning that cannot be easily encoded at the lexical-semantic level of annotation: sense interpretation may require appeal to e.g. extra-linguistic (world) knowledge which cannot be encoded/captured at the lexical-semantic level of description. We refer here to metaphors even extended to entire sequences and not limited to the single word; to words acquiring a specific sense, strictly dependent on the context, that cannot be encoded at the lexical-semantic level; or to the complexity and variety of nuances implied e.g. by a verb, according to the type of direct object co-occurring with it. Not all these shifts of meaning can or must be captured through lexical-semant annotation.

[7] This is the case of the verb *dire* (to say/to tell) which has the following synsets, among others, in IWN:
dire, enunciare, proferire (utter, mouth, etc.)
spiegare, dire (explain, tell)
dire, far sapere (tell, let it be known).

References

Antonietta Alonge, Francesca Bertagna, Nicoleti Calzolari, Adriana Roventini, Antonio Zampol (2000) *Encoding information on adjectives in lexical-semantic net for computational application.* Proceedings of the 1st NAACL Meeting, Seattle, pp 42-49.

Adam Kilgarriff (1997) *"I don't believe in wor senses"*. ITRI-97-12.

Alessandro Lenci, Nuria Bel, Federica Busa, Nicolett Calzolari, Elisabetta Gola, Monica Monachini Antoine Ogonowsky, Ivonne Peters, Wim Peters Nilda Ruimy, Marta Villegas, Antonio Zampoll (2000) *SIMPLE: A General Framework for th Development of Multilingual Lexicons*. Internationa Journal of Lexicography, XIII (4): pp. 249-263.

Simonetta Montemagni, Francesco Barsotti, Marc Battista, Nicoletta Calzolari, Ornella Corazzari Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, Rodolfo Delmonte (2000) *The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation.* LINC-2000, Luxembourg.

Simonetta Montemagni, Barsotti Francesco, Battista Marco, Calzolari Nicoletta, Corazzari Ornella, Lenci Alessandro, Zampolli Antonio, Fanciulli Francesca, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, Rodolfo Delmonte, (2000) *Building the Italian Syntactic-Semantic Treebank.* In "Building and Using Syntactically Annotated Corpora", A. Abeillé, ed., Language and Speech Series, KLUWER, Dordrecht.

Adriana Roventini, Antonietta Alonge, Nicoletta Calzolari, Bernardo Magnini, Francesca Bertagna (2000) *ItalWordNet: a large semantic database for Italian*. Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens, Greece.

Piek Vossen (ed.) (1999) *EuroWordNet General Document*, http://www.hum.uva.nl/~ewn.