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

The Italian Lexical Sample Task at SENSEVAL-3

Bernardo Magnini, Danilo Giampiccolo and Alessandro Vallin
ITC-Irst, Istituto per la Ricerca Scientifica e Tecnologica
Via Sommarive, 18 – 38050 Trento, Italy
{magnini, giampiccolo, vallin}@itc.it

Abstract

The Italian lexical sample task at SENSEVAL-3 provided a framework to evaluate supervised and semi-supervised WSD systems. This paper reports on the task preparation – which offered the opportunity to review and refine the Italian MultiWordNet – and on the results of the six participants, focussing on both the manual and automatic tagging procedures.

1 Introduction

The task consisted in automatically determining the correct meaning of a word within a given context (i.e. a short text snippet). Systems' results were compared on the one hand to those achieved by human annotators (upper bound), and on the other hand to those returned by a basic algorithm (baseline).

In the second section of this paper an overview of the task preparation is given and in the following one the main features of the participating systems are briefly outlined and the results of the evaluation exercise are presented.

In the conclusions we give an overall judgement of the outcome of the task, suggesting possible improvements for the next campaign.

2 Manual Annotation

A collection of manually labeled instances was built for three main reasons:

1. automatic evaluation (using the Scorer2 program) required a Gold Standard list of senses provided by human annotators;
2. supervised WSD systems need a labeled set of training data, that in our case was twice larger than the test set;
3. manual semantic annotation is a time-consuming activity, but SENSEVAL represents the framework to build reusable benchmark resources. Besides, manual sense tagging entails the revision of the sense inventory, whose granularity does not always satisfy annotators.

2.1 Corpus and Words Choice

The document collection from which the annotators selected the text snippets containing the lemmata to disambiguate was the *macro-balanced* section of the Meaning Italian Corpus (Bentivogli et al., 2003). This corpus is an open domain collection of newspaper articles that contains about 90 million tokens covering a time-spam of 4 years (1998-2001). The corpus was indexed in order to browse it with the Toolbox for Lexicographers (Giuliano, 2002), a concordancer that enables taggers to highlight the occurrences of a token within a context.

Two taggers chose 45 lexical entries (25 nouns, 10 adjectives and 10 verbs) according to their polysemy in the sense inventory, their polysemy in the corpus and their frequency (Edmonds, 2000). The words that had already been used at SENSEVAL-2 were avoided. Ten words were shared with the Spanish, Catalan and Basque lexical sample tasks.

Annotators were provided with a formula that indicated the number of labeled instances for each lemma[1], so they checked that the words were con-

[1] No. of labeled instances for each lemma = 75 + (15*no. of attested senses) + (7* no. of attested multiwords), where 75 is a fixed number of examples distributed over all the attested senses.

siderably frequent and polysemous before starting to tag and save the instances.
As a result, average polysemy attested in the labeled data turned out to be quite high: six senses for the nouns, six for the adjectives and seven for the verbs.

2.2 Sense Inventory and Manual Tagging

Differently from the Italian lexical sample task at SENSEVAL-2, where the instances were tagged according to ItalWordNet (Calzolari et al., 2002), this year annotators used the Italian MultiWordNet, (hereafter MWN) developed at ITC-Irst (Pianta, 2002). This lexical-semantic database includes about 42,000 lemmata and 60,000 word senses, corresponding to 34,000 synsets. Instead of distributing to participants the senses of each lemma and a limited hierarchical data structure of the semantic relations of the senses (as happened at SENSEVAL-2), the entire resource was made available. Nevertheless, none of the six participating systems, being supervised, actually needed MWN.

The annotators' task was to tag one occurrence of each selected word in all the saved instances, assigning only one sense drawn from the Italian MWN. The Toolbox for Lexicographers enabled annotators to browse the document collection and to save the relevant text snippets, while a graphical interface[2] was used to annotate the occurrences, storing them in a database. Generally, instances consisted of the sentence containing the ambiguous lemma, with a preceding and a following sentence. Nevertheless, annotators tended to save the minimal piece of information that a human would need to disambiguate the lemma, which was often shorter than three sentences.

The two annotators were involved simultaneously: firstly, each of them saved a part of the instances and tagged the occurrences, secondly they tagged the examples that had been chosen by the other one.

More importantly, they interacted with a lexicographer, who reviewed the sense inventory whenever they encountered difficulties. Sometimes there was an overlap between two or more word senses, while in other cases MWN needed to be enriched, adding new synsets, relations or definitions. All the 45 lexical entries we considered were thoroughly reviewed, so that word senses were as clear as possible to the annotators. On the one hand, the revision of MWN made manual tagging easier, while on the other hand it led to a high Inter Tagger Agreement (that ranged between 73 and 99 per cent), consequently reflected in the K statistics (that ranged between 0.68 and 0.99).

Table 1 below summarizes the results of the manual tagging.

	Average polysemy in MWN	Average polysemy in the labeled set	I.T.A. Average K	# training examples	# test examples
25 nouns	10	6	0.9	2835	1343
10 adjectives	8	6	0.89	1111	524
10 verbs	9	7	0.89	1199	572

Table 1. Manual Annotation Results

Once the instances had been collected and tagged by both the annotators, we asked them to discuss the examples about which they disagreed and to find a definitive meaning for them.

Since the annotators built the corpus while tagging, they tended to choose occurrences whose meaning was immediately straightforward, avoiding problematic cases. As a consequence, the ITA turned out to be so high and the distribution of the senses in the labeled data set did not reflect the actual frequency in the Italian language, which may have affected the systems' performance.

Annotators assigned different senses to 674 instances over a total of 7584 labeled examples. Generally, disagreement depended on trivial mistakes, and in most cases one of the two assigned meanings was chosen as the final one. Nevertheless, in 46 cases the third and last annotation was different from the previous two, which could demonstrate that a few word senses were not completely straightforward even after the revision of the sense inventory.

For example, the following instance for the lemma "vertice" (vertex, acme, peak) was annotated in three different ways:

*La struttura lavorativa – spiega Grandi – ha un carattere paramilitare. Al **vertice** della piramide c'è il direttore, poi i manager, quelli con la cravatta e la camicia a mezze maniche.*

Annotator 1 tagged with sense 2 (Factotum, "the highest point of something"), while annotator 2 decided for sense 4 (Geometry, "the point of in-

[2] This tool was designed and developed by Christian Girardi at ITC-Irst, Trento, Italy.

tersection of lines or the point opposite the base of a figure") because the text refers to the vertex of a pyramid. Actually, the snippet reported this abstract image to describe the structure of an enterprise, so in the end the two taggers opted for sense 5 (Administration, "the group of the executives of a corporation"). Therefore, subjectivity in manual tagging was considerably reduced by adjusting the sense repository and selecting manually each single instance, but it could not be eliminated.

3 Automatic Annotation

We provided participants with three data sets: labeled training data (twice larger than the test set), unlabeled training data (about 10 times the labeled instances) and test data. In order to facilitate participation, we PoS-tagged the labeled data sets using an Italian version of the TnT PoS-tagger (Brants, 2000), trained on the Elsnet corpus.

3.1 Participants' results

Three groups participated in the Italian lexical sample task, testing six systems: two developed by ITC-Irst - Italy - (*IRST-Kernels* and *IRST-Ties*), three by Swarthmore College - U.S.A. - (*swat-hk-italian*, *Italian-swat_hk-bo* and *swat-italian*) and one by UNED - Spain.

Table 2 below reports the participants' results, sorted by F-measure.

system	precision	recall	attempted	F-measure
IRST-Kernels	0.531	0.531	100%	0.531
swat-hk-italian	0.515	0.515	100%	0.515
UNED	0.498	0.498	100%	0.498
italian-swat_hk-bo	0.483	0.483	100%	0.483
swat-italian	0.465	0.465	100%	0.465
IRST-Ties	0.552	0.309	55.92%	0.396
baseline	0.183	0.183	100%	0.183

Table 2. Automatic Annotation Results (fine-grained score)

The baseline results were obtained running a simple algorithm that assigned to the instances of the test set the most frequent sense of each lemma in the training set. All the systems outperformed the baseline and obtained similar results. Compared to the baseline of the other Lexical Sample tasks, ours is much lower because we interpreted the formula described above (see footnote 1), and tagged the same number of instances for all the senses of each lemma disregarding their frequency in the document collection. As a result, the distribution of the examples over the attested senses did not reflect the one in natural language, which may have affected the systems' performance.

While at SENSEVAL-2 test set senses were clustered in order to compute mixed- and coarse-grained scores, this year we decided to return just the fine-grained measure, where an automatically tagged instance is correct only if the sense corresponds to the one assigned by humans, and wrong otherwise (i.e. one-to-one mapping).

There are different sense clustering methods, but grouping meanings according to some sort of similarity is always an arbitrary decision. We intended to calculate a domain-based coarse-grained score, where word senses were clustered according to the domain information provided in WordNet Domains (Magnini and Cavaglià, 2000). Unfortunately, this approach would have been significant with nouns, but not with adjectives and verbs, that belong mostly to the generic Factotum domain, so we discarded the idea.

All the six participating systems were supervised, which means they all used the training data set and no one utilized either unlabelled instances or the lexical database. UNED used also SemCor as an additional source of training examples.

IRST-Kernels system exploited Kernel methods for pattern abstraction and combination of different knowledge sources, in particular paradigmatic and syntagmatic information, and achieved the best F-measure score.

IRST-Ties, a generalized pattern abstraction system originally developed for Information Extraction tasks and mainly based on the boosted wrapper induction algorithm, used only lemma and POS as features. Proposed as a "baseline" system to discover syntagmatic patterns, it obtained a quite low recall (about 55 per cent), which affected the F-measure, but proved to be the most precise system.

Swarthmore College wrote three supervised classifiers: a clustering system based on cosine similarity, a decision list system and a naive bayes classifier. Besides, Swarthmore group took advantage of two systems developed at the Hong Kong Polytechnic University: a maximum entropy classifier and system which used boosting (*Italian-swat_hk-bo)*. The run *swat-hk-italian* joined all the

five classifiers according to a simple majority-vote scheme, while *swat-hk-italian* did the same using only the three classifiers developed in Swarthmore.

The system presented by the UNED group employed similarity as a learning paradigm, considering the co-occurrence of different nouns and adjectives.

3.2 General Remarks on Task Complexity

As we mentioned above, the 45 words for the Italian lexical sample task were chosen according to their polysemy and frequency. We addressed difficult words, that had at least 5 senses in MWN.

Actually, polysemy does not seem to be directly related to systems' results (Calzolari, 2002), in fact the average F-measure of our six runs for the nouns (0.512) was higher than for adjectives (0.472) and verbs (0.448), although the former had more attested senses in the labeled data.

Complexity in returning the correct sense seems to depend on the blurred distinction between similar meanings rather than on the number of senses themselves. If we consider the nouns "attacco" (attack) and "esecuzione" (performance, execution), for which the systems obtained the worst and one of the best average results respectively, we notice that the 4 attested senses of "esecuzione" were clearly distinguished and referred to different domains (Factotum, Art, Law and Politics), while the 6 attested senses of "attacco" were more subtly defined. Senses 2, 7 and 11 were very difficult to discriminate and often appeared in metaphorical contexts. Senses 5 and 6, for their part, belong to the Sport domain and are not always easy to distinguish.

4 Conclusions

The results of the six systems participating in the evaluation exercise showed some improvements compared to the average performance at SENSEVAL-2, though data sets and sense repositories were considerably different.

We are pleased with the successful outcome of the experiments in terms of participation, although regrettably no system exploited the unlabeled training set, which was intended to offer a less time-consuming resource. On the other hand, the labeled instances that have been collected represent a useful and reusable benchmark.

As a final remark we think it could be interesting to consider the actual distribution of word senses in Italian corpora in collecting the examples for the next campaign.

Acknowledgements

We would like to thank Christian Girardi and Oleksandr Vagin for their technical support; Claudio Giuliano and the Ladin Cultural Centre for the use of their Toolbox for Lexicographers; Pamela Forner, Daniela Andreatta and Elisabetta Fauri for the revision of the Italian MWN and on the semantic annotation of the examples; and Luisa Bentivogli and Emanuele Pianta for their precious suggestions during the manual annotation.

References

Luisa Bentivogli, Christian Girardi and Emanuele Pianta. 2003. The MEANING Italian Corpus. In *Proceedings of the Corpus Linguistics 2003 conference*, Lancaster, UK: 103-112.

Francesca Bertagna, Claudia Soria and Nicoletta Calzolari. 2001. The Italian Lexical Sample Task. In *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France: 29-32.

Thorsten Brants. 2000. TnT - a Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA: 224-231.

Nicoletta Calzolari, Claudia Soria, Francesca Bertagna and Francesco Barsotti. 2002. Evaluating lexical resources using SENSEVAL. *Natural Language Engineering*, 8(4): 375-390.

Philip Edmonds. 2000. Designing a task for SENSEVAL-2. (http://www.sle.sharp.co.uk/SENSEVAL2/archive/index.htm)

Claudio Giuliano. 2002. A Toolbox for Lexicographers In *Proceedings of the tenth EURALEX International Congress*, Copenhagen, Denmark: 113-118.

Bernardo Magnini and Gabriela Cavaglià. 2000. Integrating Subject Field Codes into WordNet. In *Proceedings of LREC-2000*, Athens, Greece: 1413-1418.

Emanuele Pianta, Luisa Bentivogli and Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, Mysore, India: 293-302.