Evalita 2011: Description and Results of the SuperSense Tagging Task

Stefano Dei Rossi, Giulia Di Pietro, Maria Simi

Università di Pisa, Dipartimento di Informatica,
Largo B. Pontecovo 3, 56127 Pisa, Italy
{deirossi, simi, evalita2011}@di.unipi.it

Abstract. SuperSense tagging (SST) is a Natural Language Processing task that consists in annotating each significant entity in a text, like nouns, verbs, adjectives and adverbs, within a general semantic taxonomy defined by the WordNet lexicographer classes (called SuperSenses). SST can be considered as a task half-way between Named-Entity Recognition (NER) and Word Sense Disambiguation (WSD): it is an extension of NER, since it uses a larger set of semantic categories, and it is an easier and more practical task with respect to WSD, that deals with very specific senses. We will report on the organization and results of the Evalita 2011 SuperSense Tagging task.

Keywords: SuperSense Tagging, SST, Named Entity Recognition, NER, Word Sense Disambiguation, WSD

1 Motivation

SuperSense tagging (SST) is a Natural Language Processing task that consists in annotating each significant entity in a text, like nouns, verbs, adjectives and adverbs, within a general semantic taxonomy defined by the WordNet lexicographer classes (called SuperSenses) [1]. SST can be considered as a task half-way between Named-Entity Recognition (NER) and Word Sense Disambiguation (WSD) since it is an extension of NER, as it uses a larger set of semantic categories, and it is an easier and more practical task with respect to WSD, that deals with very specific senses. SST can therefore be of practical value in a number of NLP tasks involving world knowledge such as semantic information retrieval, question answering and information extraction.

A preliminary version of the corpus for SuperSense tagging, called ISST-SST, was created starting from the Italian Syntactic-Semantic Treebank (ISST) [2] as part of the project SemaWiki (Text Analytics and Natural Language processing - Tanl) [4], a collaboration between the University of Pisa and the Institute for Computational Linguistics of CNR. The details of its construction were presented in [3].

The Evalita 2011 challenge was an opportunity to complete, revise and extend ISST-SST.

2 Definition of the Task

The goal of the Evalita 2001 task is to predict an appropriate SuperSense for each token or multiword expression. Modal and support verbs are not annotated since they do not entail any semantics.

The 45 SuperSense categories (3 used for adjectives, 25 for nouns, 15 for verbs and one for adverbs) are shown in the following table. Some more detail is provided in the guidelines for the task and in [3].

Table 1. WordNet lexicographer classes (SuperSenses)

Id	SuperSense	Description
00	adj.all	all adjective clusters, used for all simple adjectives
01	adj.pert	relational adjectives (pertainyms), adjectives that are related with nouns
02	adv.all	all adverb
03	noun.Tops	unique beginner for nouns, nouns that appear at top level
04	noun.act	nouns denoting acts or actions
05	noun.animal	nouns denoting animals
06	noun.artifact	nouns denoting man-made objects
07	noun.attribute	nouns denoting attributes of people and objects
08	noun.body	nouns denoting body parts
09	noun.cognition	nouns denoting cognitive processes and contents
10	noun.communication	nouns denoting communicative processes and contents
11	noun.event	nouns denoting natural events
12	noun.feeling	nouns denoting feelings and emotions
13	noun.food	nouns denoting foods and drinks
14	noun.group	nouns denoting groupings of people or objects
15	noun.location	nouns denoting spatial position
16	noun.motive	nouns denoting goals
17	noun.object	nouns denoting natural objects (not man-made)
18	noun.person	nouns denoting people
19	noun.phenomenon	nouns denoting natural phenomena
20	noun.plant	nouns denoting plants
21	noun.possession	nouns denoting possession and transfer of possession
22	noun.process	nouns denoting natural processes
23	noun.quantity	nouns denoting quantities and units of measure
24	noun.relation	nouns denoting relations between people or things or ideas
25	noun.shape	nouns denoting two and three dimensional shapes
26	noun.state	nouns denoting stable states of affairs
27	noun.substance	nouns denoting substances
28	noun.time	nouns denoting time and temporal relations
29	verb.body	verbs of grooming, dressing and bodily care
30	verb.change	verbs of size, temperature change, intensifying, etc.
31	verb.cognition	verbs of thinking, judging, analyzing, doubting

32	verb.communication	verbs of telling, asking, ordering, singing
33	verb.competition	verbs of fighting, athletic activities
34	verb.consumption	verbs of eating and drinking
35	verb.contact	verbs of touching, hitting, tying, digging
36	verb.creation	verbs of sewing, baking, painting, performing
37	verb.emotion	verbs of feeling
38	verb.motion	verbs of walking, flying, swimming
39	verb.perception	verbs of seeing, hearing, feeling
40	verb.possession	verbs of buying, selling, owning
41	verb.social	verbs of political and social activities and events
42	verb.stative	verbs of being, having, spatial relations
43	verb.weather	verbs of raining, snowing, thawing, thundering
44	adj.ppl	participial adjectives

Two subtasks were organized:

- *Closed subtask***.** The aim of the closed subtask was to measure the accuracy in SuperSense tagging, using only the corpus provided for training
- *Open subtask***.** In the open subtask participants could use any external resource in addition to the corpus provided for training; for example, instances of WordNet as well as other lexical or semantic resources.

3 Dataset

ISST-SST (about 300,000 tokens) was first made available for research purposes in 2010, and was completely revised for the Evalita 2011 task. The work consisted in completing the tagging (24,000 tokens were not tagged in the previous version) and by revising the tagging strategy for multi-word expressions. For instance expressions such as “Croce Rossa”, “Fiona May” and “10 dicembre 1975” are now considered as single entities.

A portion of about 276,000 tokens from the revised corpus was used in the Evalita 2011 task for training and development. The evaluation was performed on a smaller corpus obtained from a held-out portion of ISST-SST (about 30,000 tokens) and a brand new portion of the Italian Wikipedia (about 20,000 additional tokens), annotated and manually revised for this task.

Data adhere to the following rules:

1. Characters are UTF-8 encoded (Unicode).
2. Data files are organized in documents.
3. Each document contains sentences separated by an empty line.
4. A sentence consists of a sequence of tokens, one token per line.
5. A token consists of four fields (separated by tabs characters): FORM, LEMMA, PoS, SuperSense. All these fields were produced with automatic tools and manually revised.

6. SST tags can span several tokens and use the IOB2 notation: labels are prefixed with "B" for begin, "I" for inside, and "O", outside any label.

Example of Annotation.

```
Un           un           RIms     O
incendio     incendio     Sms      B-noun.event
,            ,            FF       O
che          che          PRnn     O
si           si           PC3nn    O
sarebbe      essere       VAd3s    O
sviluppato   sviluppare   Vpsms    B-verb.creation
per          per          E        O
cause        causa        Sfp      B-noun.motive
accidentali  accidentale  Anp      B-adj.all
,            ,            FF       O
ha           avere        VAip3s   O
gravemente   gravemente    B        B-adv.all
danneggiato  danneggiare   Vpsms    B-verb.change
a            a            E        O
Fiano        fiano        SP       B-noun.location
,            ,            FF       O
uno          uno          RIms     O
chalet       chalet       Smn      B-noun.artifact
di           di           E        O
proprietà    proprietà    Sfn      B-noun.possession
di           di           E        O
Umberto      umberto      SP       B-noun.person
Agnelli      agnelli      SP       I-noun.person
```

4 Evaluation Measures

The evaluation metrics are quite standard:

- *Tagging accuracy*, i.e. the percentage of correctly classified tokens with respect to the total number of tokens;
- *F1-measure*, the weighted harmonic mean of *precision* and *recall.*

5 Participation Results

Only two teams submitted runs for the SuperSense Tagging Task: the University of Pisa (UNIPI - Simi et al.) and the University of Bari (UNIBA – Basile).

The UniPI team participated only to the closed subtask with a system based on a Maximum Entropy classifier, for learning how to chunk texts, and a dynamic programming algorithm, in order to select sequences of tags with the highest

probability. The tagger extracts three kinds of features: attributes features, related to the attributes of surrounding tokens; local features, i.e. features related to the *shape* of the current word and its context; global features that are properties holding at the document level. The four runs, called **run [1-4]**, were created using the same set of local features, together with different permutations of the attributes features and a different number of iteration. In runs 3 and 4 a more specialized set of local features were added, with the aim to improve the performance on a subset of the SuperSenses.

The UniBA team participated to both subtasks with two different systems, both based on Support Vector Machines classifiers. In two of the open task runs some features provided by a semantic *WordSpace* were used with the aim of solving the data sparseness problem. The core idea behind the *WordSpace* is that words and concepts are represented by points in a mathematical space, and concepts with similar or related meanings are near to one another in that space. The meaning of a word is determined by the rules of its usage in the context of ordinary and concrete language behavior, hence, words are semantically similar if they share contexts.

The runs were created using the following features:

- **uniba SST Closed yc:** form, lemma and PoS of the current, previous and following token (baseline features);
- **uniba SST Open yo:** baseline features plus the SuperSense assigned to the most frequent sense of the current word (computed according to sense frequency in MultiSemCor), current word upper-case, the grammatical conjugation of the word and the coarse-grained PoS;
- **uniba SST Open SVMcat:** distributional information about word contexts built using Wikipedia categories;
- **uniba SST Open SVMterm:** distributional information about word contexts built on Wikipedia pages contexts.

Systems Results. The results of the systems are summarized in the following tables:

Table 2. Closed task

	Accuracy	Precision	Recall	F1 test	F1 ISST	F1 Wiki
UniPI - run 3	**88.50%**	**76.82%**	**79.76%**	**78.27**	**78.23**	**78.36**
UniPI - run 2	88.34%	76.69%	79.38%	78.01	78.33	77.28
UniPI - run 1	88.30%	76.64%	79.33%	77.96	78.20	77.42
UniPI - run 4	88.27%	76.48%	79.29%	77.86	78.15	77.20
UniBA - yc	86.96%	74.85%	75.83%	75.34	76.29	73.38

Table 3. Open task

	Accuracy	Precision	Recall	F1	F1 ISST	F1 Wiki
UniBA - SVMcat	**88.77%**	**77.19%**	**80.20%**	**78.66**	**79.69**	**76.29**
UniBA – SVMterm	88.64%	77.00%	79.98%	78.46	79.59	75.86
UniBA - yo	88.22%	77.28%	78.18%	77.73	78.10	76.86

6 Discussion

The best performances obtained by the systems of the two teams are very good and very close. For the Closed Task the best system is from UNIPI, slightly outperformed by the best system from UNIBA in the Open Task. A closer look at the differences on key categories is presented in Fig 1.

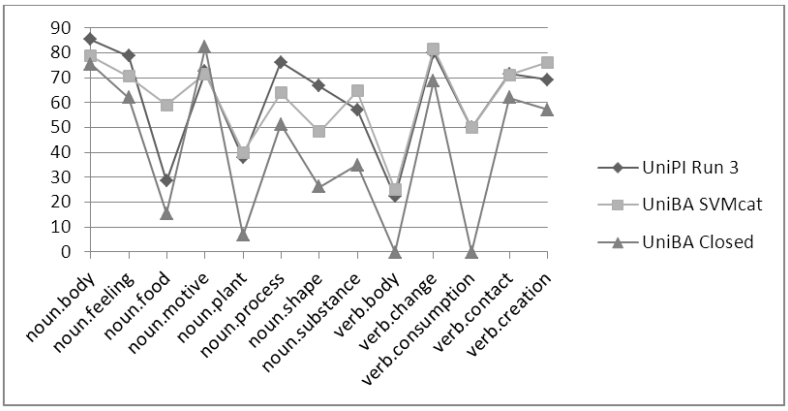

Fig. 1. Comparison of three systems on the 13 categories where the performance difference between systems is higher

System results in Table 2 and 3 also highlight the different performances on the two subparts of the test set: the one in the same domain as the training corpus (ISST), and the one derived from articles in Wikipedia (Wiki). The models learned on the ISST-SST training set seem to be able to cope effectively with a different domain, without any specific adaptation strategy. This is especially true for the UNIPI system, while in the case of UNIBA we observe a slight decay in performance (on the order of 2-3 points of F1).

Acknowledgements. We thank G. Attardi, A. Lenci, S. Montemagni, and the other members of the SemaWiki project team for the work which made this task possible.

References

1. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
2. Montemagni, S. et al.: Building the Italian Syntactic-Semantic Treebank. In: Abeillé (ed.), Building and using Parsed Corpora, Language and Speech series, pp. 189–210, Kluwer, Dordrecht (2003)
3. Attardi, G., Dei Rossi, S., Di Pietro, G., Lenci, A., Montemagni, S., Simi, M.: A Resource and Tool for SuperSense Tagging of Italian Texts. In: 7th Language Resources and Evaluation Conference, LREC 2010, Malta (2010)
4. Tanl (Text Analytics and Natural Language processing). Project: Analisi di Testi per il Semantic Web e il Question Answering, http://medialab.di.unipi.it/wiki/SemaWiki (2009)