Supersense Tagger for Italian

Davide Picca, Alfio Massimiliano Gliozzo, Massimiliano Ciaramita

University of Lausanne - CH-1015 Lausanne - Switzerland
davide.picca@unil.ch
Laboratory for Applied Ontology, ISTC-CNR, Rome, Italy
alfio.gliozzo@istc.cnr.it
Yahoo! Research Barcelona Ocata 1 08003 Barcelona - Spain
massi@yahoo-inc.com

Abstract

In this paper we present the procedure we followed to develop the Italian Super Sense Tagger. In particular, we adapted the English SuperSense Tagger to the Italian Language by exploiting a parallel sense labeled corpus for training. As for English, the Italian tagger uses a fixed set of 26 semantic labels, called supersenses, achieving a slightly lower accuracy due to the lower quality of the Italian training data. Both taggers accomplish the same task of identifying entities and concepts belonging to a common set of ontological types. This parallelism allows us to define effective methodologies for a broad range of cross-language knowledge acquisition tasks

1. Introduction

Developed in the Information Extraction field, Named Entity Recognition (NER) is a basic task in Natural Language Processing. NER, originally exploited for business activities (Grishman and Sundheim, 1996), has been extended beyond this field. In particular, NER can be a useful step for broad-coverage ontology engineering. For example, named entity categories could be used for ontology population and organization. New pertinent categories, in addition to the classical ones, are likely to be useful in order to build a taxonomic hierarchy. The first problem is the definition of the categories. Even for small sets of categories their definition tends to be controversial. As well-discussed in (Sekine, 2004), the outstanding issue to find good categories for NER refers to the problem of categorizing the world into semantic categories, and finding the right category for each word (Sekine, 2004). In practice one limitation of NER is the fact that traditional categories (e.g., person, location, and organization) are too few and generic.

One interesting alternative to traditional NER categories are the most general, or top-level, categories defined by WordNet. WordNet as been organized according to psycholinguistic theories on the principles governing lexical memory. As an example, several psycho-linguistic experiments discussed in (Miller, 1990) suggest correlations between reaction times and the hierarchical structural of the lexicon. Thus the broadest WordNet's categories can serve as a principled basis for a set of categories which exhaustively covers, at least as a first rough approximation, all possible concepts occurring in a sentence. An additional advantage of such categories is that, in principle, they should be categories which are shared across different languages. Thus, semantic annotations of this kind could be used for multilingual inference in several language tasks; e.g., information retrieval or machine translation.

To this aim, (Ciaramita and Johnson, 2003) developed a SuperSense Tagging (SST) technology for English, demonstrating that reasonably high accuracy in tagging can be obtained even in open domain contexts. This technology has been also adopted for Ontology Learning (Picca et al., May 2007), as the top level WordNet SuperSenses cover almost any high level ontological type of interest in ontology design. Section 2. describes the main features of the English SST.

In this paper we investigate the problem of developing a tagger based on WordNet semantic categories for Italian. The basic idea is that, being the WordNet supersenses inherently multilingual, the SST technology can be adopted for multilingual ontology learning problems. To this aim, we ported the SST technology to Italian, by training the supervised learning algorithm at the basis of the English distribution of the SST on an Italian sense tagged corpus, called MultiSemCor (Bentivogli et al., 2004). The procedure adopted to this aim is described in Section 3. In section 4. we evaluate the quality of the so generated Italian tagger. The results are comparable to those obtained for English, however the noise introduced by the steps that are necessary in order to generate the training data is considerable and further research is needed to improve them. Finally, Section 5. summarizes the achieved results proposing some idea to apply both the English and Italian SSTs for multilingual knowledge acquisition problems.

2. The English SuperSense Tagger

WordNet (Fellbaum, 1998) defines 45 lexicographer's categories, also called *supersenses* (Ciaramita and Johnson, 2003), used by lexicographers to provide an initial broad classification for the lexicon entries[1]. Although simplistic in many ways, the supersense ontology has several attractive features for NLP purposes. First, concepts, although fairly general, are easily recognizable. Secondly, the small number of classes makes it possible to implement state of the art methods, such as sequence taggers, to annotate text with supersenses. Finally, similar word senses tend to be merged together reducing ambiguity. Hence, while the noun *folk* has four fine-grained senses, at the supersense level it only has two, as illustrated below:

1. people in general (noun.group)

[1] We used the WordNet version 2.0 for all the experiments in the paper.

2. a social division of (usually preliterate) people (noun.group)

3. people descended from a common ancestor (noun.group)

4. the traditional and typically anonymous music that is an expression of the life of people in a community (noun.communication)

Previous work has showed that supersenses can be useful in lexical acquisition to provide a first guess at the meaning of novel words (Ciaramita and Johnson, 2003), and in syntactic parse re-ranking, to define latent semantic features (Picca et al., May 2007) (Koo and Collins, 2005). Using the Semcor corpus, a fraction of the Brown corpus annotated with WordNet word senses, a SST has been implemented (Ciaramita and Altun, 2006) which can be used for annotating large collections of English text [2]. The SST implements a Hidden Markov Model, trained with the perceptron algorithm introduced in (Collins, 2002). Perceptron sequence learning provides an excellent trade-off accuracy/performance, sometimes outperforming more complex models such as CFR (Nguyen and Guo, 2007). The tagset used by the tagger defines 26 supersense labels for nouns and 15 supersense labels for verbs. The basic feature set includes:

- *word* = lower-cased form of each token for the current position i and in addition for i-1 and i+1
- *sh* = shape of the token as a simple regular expression-like representation
- *pos* = POS of i, i-1 and i+1
- *sb*= bi- and tri-grams of characters of the suffix of word_i
- *pr*= bi- and tri-grams of characters of the prefix of word_i
- *rp* = coarse relative position of word_i, rp=begin if i = 0, rp=end if i = —sentence—-1, sb=mid otherwise
- *kf* = constant features on each token for regularization purposes

In addition to this set, the Most Frequent sense in WordNet is also provided as an additional feature for the English SST, exploiting the fact that English Word Senses in WordNet are ordered by frequency.

The tagger outputs Named Entity information, but also covers other relevant categories and attempts lexical disambiguation at the supersense level. The following is a sample output of the tagger:

```
Guns_B-noun.group and_I-noun.group
Roses_I-noun.group
plays_B-verb.communication at_O
the_O stadium_B-noun.location
```

Compared to other semantic tagsets, supersenses have the advantage of being designed to cover all possible open class words. Thus, in principle, there is a supersense category for each word, known or novel. Additionally, no distinction is made between proper and common nouns, whereas standard Named Entity Recognition systems tends to be biased towards the former.

3. Porting the SST technology to Italian

In order to fulfill our research direction in multilingual ontology learning, we ported the English SST to Italian. To this aim, we need the following resources for the Italian language:

1. An Italian POS tagger
2. An Italian Sense Tagged corpus for training, where words are tagged with WordNet super-senses

As a PoS tagger, we adopted the Evalita Tagset (Ciaramita and Atserias, 2007), a tool for annotating text with part-of-speech and lemma information.

As a source of sense tagged data, we adopted MultiSemCor (Bentivogli et al., 2004), an Italian corpus composed of 116 texts which are the translation of their corresponding English texts in SemCor. This resource has been developed by manually translating the English texts to Italian. Then, the so generated parallel corpus has been automatically aligned at the Word Level. Finally, sense labels have been automatically transferred from the English words to their Italian translations.

The sense labels adopted in the Italian part of MultiSemCor (Bentivogli et al., 2004) have been extracted by Multi WordNet [3]. It is a multilingual computational lexicon, conceived to be strictly aligned with the Princeton WordNet. The available languages are Italian, Spanish, Hebrew and Romanian. In our experiment we used the English and the Italian components. The last version of the Italian WordNet contains around 58,000 Italian word senses and 41,500 lemmas organized into 32,700 synsets aligned with WordNet English synsets. The Italian synsets are created in correspondence with the Princeton WordNet synsets, whenever possible, and semantic relations are imported from the corresponding English synsets. This implies that the synset index structure is the same for the two languages.

The full alignment between the English and the Italian WordNet is guaranteed by the fact that both resources adopts the same synsetIDs to refer to concepts. This nice feature allowed us to infer the correct super-sense for each Italian sense by simply looking at the English structure. In this way, we assign exactly the same ontological types to both Italian and English terms, thus obtaining an Italian corpus tagged by its supersenses as shown below:

```
La ART 0
contea NN B-noun.location
di PREP I-noun.location
Fulton NN_P I-noun.location
deve V_MOD 0
ricevere V_GVRB B-verb.possession
```

[2] The tagger is publicly available at: http://sourceforge.net/projects/supersensetag/.

[3] Available at *http://multi WordNet.itc.it*.

```
una ART 0
porzione NN B-noun.act
di PREP 0
...
```

As a feature set, we adopted the default configuration described in Section 2., avoiding the use of the First Sense feature, as the sense order is not representative for the Italian part of Multi WordNet.

Then, we trained the SST engine described in Section 2. (obtained from the original distribution) in the corpus generated so far, and we optimized the required parameters by adopting a cross validation technique. As for the English settings (Ciaramita and Johnson, 2003), the best results have been obtained by setting 50 trials and 10 epochs to train the perceptron algorithm.

Results and error analysis are presented in the following section.

4. Evaluation

We evaluated the performances of the Italian SST generated so far by adopting a n-fold cross validation strategy on the Italian Corpus adopted for training. Results for verbs and nouns are illustrated in Table 1, reporting precision, recall and F1 for any SuperSense. Even if the micro figure obtained is sensibly lower than the corresponding value for English (which is around 0.77, evaluated with the same procedure on the English SemCor as reported in (Ciaramita and Altun, 2006)), the results are really encouraging, achieving a micro F1 of 0.63%. If we cast a deeper glance at the Table 1, we can clearly notice that for some category the F1 is exceptionally high. Some of those best categorized categories are really essential for ontology learning. For example, important labels as *noun.person*, *noun.body* or *noun.time* achieve results higher than 70%.

On the other hand, the Italian tagger achieved lower performances if compared with the English one. It can be explained by (i) the lower quality of the training resource, (ii) the lower quantity of training data and (iii) the unavailability of the first sense info.

Regarding the first point, it is worthwhile to remark that even if the quality of transfer developed by (Bentivogli et al., 2004) is high, many incorrect sense transfers (around 14%) can be found. So our work suffers of the same faults, inherited by the automatic alignment. For example, we report here the most relevant errors we faced with during the preprocessing step. One of the most important error that has badly influenced the training set especially for multiword recognition is represented by the case in which the translation equivalent is indeed a cross-language synonym of the source expression but it is not a lexical unit. It occurs when a language expresses a concept with a lexical unit whereas the other language expresses the same concept with a free combination of words (for instance *occhiali da sole* annotated with the sense of *sunglasses*).

Regarding the second problem, we noticed that the quantity of sense labeled words adopted for English is higher than 200,000, whereas the amount of Italian tokens adopted is around 92,000. Therefore, the amount of Italian training data is sensibly lower, explaining the lower performances.

SuperSense	**Recall**	**Precision**	**F1**
noun.act	0.573663	0.569004	0.571317
noun.animal	0.607407	0.642636	0.624153
noun.artifact	0.675274	0.645451	0.660014
noun.attribute	0.558363	0.56179	0.560015
noun.body	0.730303	0.690676	0.709899
noun.cognition	0.610378	0.539244	0.57261
noun.communication	0.577329	0.567507	0.572373
noun.event	0.340824	0.510737	0.408749
noun.feeling	0.376543	0.469672	0.417819
noun.food	0.561404	0.613426	0.585843
noun.group	0.590812	0.628444	0.609028
noun.location	0.570652	0.528592	0.548784
noun.motive	0.625	0.661376	0.640523
noun.object	0.595238	0.582053	0.588396
noun.other	0.380952	0.457912	0.415726
noun.person	0.822983	0.755092	0.787559
noun.phenomenon	0.654971	0.629284	0.641861
noun.plant	0.6875	0.516667	0.589669
noun.possession	0.67284	0.619977	0.645076
noun.process	0.60	0.5625	0.580645
noun.quantity	0.54902	0.625706	0.584851
noun.relation	0.446541	0.425412	0.43569
noun.shape	0.363636	0.377778	0.36991
noun.state	0.557724	0.569854	0.563699
noun.substance	0.596154	0.607743	0.601854
noun.time	0.796636	0.742201	0.768447
verb.body	0.396552	0.469363	0.429808
verb.change	0.467014	0.512439	0.488639
verb.cognition	0.627756	0.599852	0.613397
verb.communication	0.629524	0.630734	0.630123
verb.competition	0.287356	0.356884	0.318336
verb.consumption	0.592593	0.542308	0.566316
verb.contact	0.35958	0.429483	0.391298
verb.creation	0.393258	0.426686	0.409158
verb.emotion	0.453552	0.506173	0.478411
verb.motion	0.433712	0.395141	0.413474
verb.perception	0.529551	0.54509	0.537189
verb.possession	0.526205	0.472816	0.49806
verb.social	0.381481	0.343345	0.361378
verb.stative	0.656296	0.674961	0.665495
verb.weather	0.557724	0.413571	0.566349
MICRO	0.635731	0.622552	0.629072

Table 1: Recall, Precision and F1 for each SuperSense

Moreover, the italian SST lacks in one of the most important feature used for the English SST, first sense heuristics. In fact, for the Italian language, the first sense baseline cannot be estimated by simply looking at the first sense in WordNet, since the order of the Italian WordNet does not reflect the frequency of senses. Therefore, we did not estimate this baseline for the Italian SST, in contrast to what has been done for the English SST.

Since in an ontology learning task the precision of categorization has to be more reliable than recall, we also reported these results in Figure 1 for all nouns. Precision is relatively higher for concrete types, such as *person*, *body* and *artifact*. In these cases, we get a precision of more than 70%. Concerning verb categories, the results are less accurate. This phenomena can be explained considering the fact that verbs are much more ambiguous than nouns.

5. Conclusion and future work

In this paper we presented a new Italian SuperSense Tagger able to recognize named entities and concepts in texts achieving reasonably high accuracy, even if much lower than the English counterpart. Anyhow, the achieved precision is reasonably high for the tagger to be applied in knowledge acquisition tasks.

These results are encouraging and this research deserves further investigations. First of all we are going to develop automatic techniques based on parallel corpora to develop SST for other languages, such as German and French, without exploiting any labeled data. Secondly, combing this tagger with the English one already developed, we offer a new multilingual tool, covering an higher spectrum of categories than traditional Named Entity Recognition systems. Being the category set totally aligned among languages, the tool can be profitably used as a preprocessing step for bilingual dictionary induction, multilingual ontology learning, and so on. Another direction we are following is the development of a new generation SST which is able to distinguish between concepts and instances of the same type. Finally, we are going to develop a WEB service able to extract terminology belonging to different supersenses from the analysis of corpora and WEB pages in multiple languages.

Acknowledgments

Alfio Massimiliano Gliozzo has been supported by the BONy project, financed by the Education and culture DG of the EU, grant agreement N 135263-2007-IT-KA3-KA3MP, under the Lifelong Learning Programme 2007 managed by EACEA.

6. References

Luisa Bentivogli, Pamela Forner, and Emanuele Pianta. 2004. Evaluating cross-language annotation transfer in the multisemcor corpus. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 364, Morristown, NJ, USA. Association for Computational Linguistics.

M. Ciaramita and Y. Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of EMNLP-06*, pages 594–602, Sydney, Australia.

M. Ciaramita and J. Atserias. 2007. Pos tagging with a named entity tagger. *Intelligenza Artificiale*, 4:28–29.

M. Ciaramita and M. Johnson. 2003. Supersense tagging of unknown nouns in wordnet. In *Proceedings of EMNLP-03*, pages 168–175, Sapporo, Japan.

M. Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP-02*.

C. Fellbaum. 1998. *WordNet. An Electronic Lexical Database*. MIT Press.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics*, pages 466–471, Morristown, NJ, USA. Association for Computational Linguistics.

T. Koo and M. Collins. 2005. Hidden-variable models for discriminative reranking. In *Proceedings of EMNLP-05*, Vancouver, Canada.

G. A. Miller. 1990. Nouns in wordnet: a lexical inheritance system,. *International Journal of Lexicography*, 3(4):245–264.

Nam Nguyen and Yunsong Guo. 2007. Comparison of sequence labeling algorithms and extensions. In *Proceedings of ICML 2007*, pages 681–688.

Davide Picca, Alfio Gliozzo, and Massimiliano Ciaramita. May 2007. Semantic domains and supersens tagging for domain-specific ontology learning. In *proceedings RIAO 2007*.

S Sekine. 2004. Named entity: History and future.

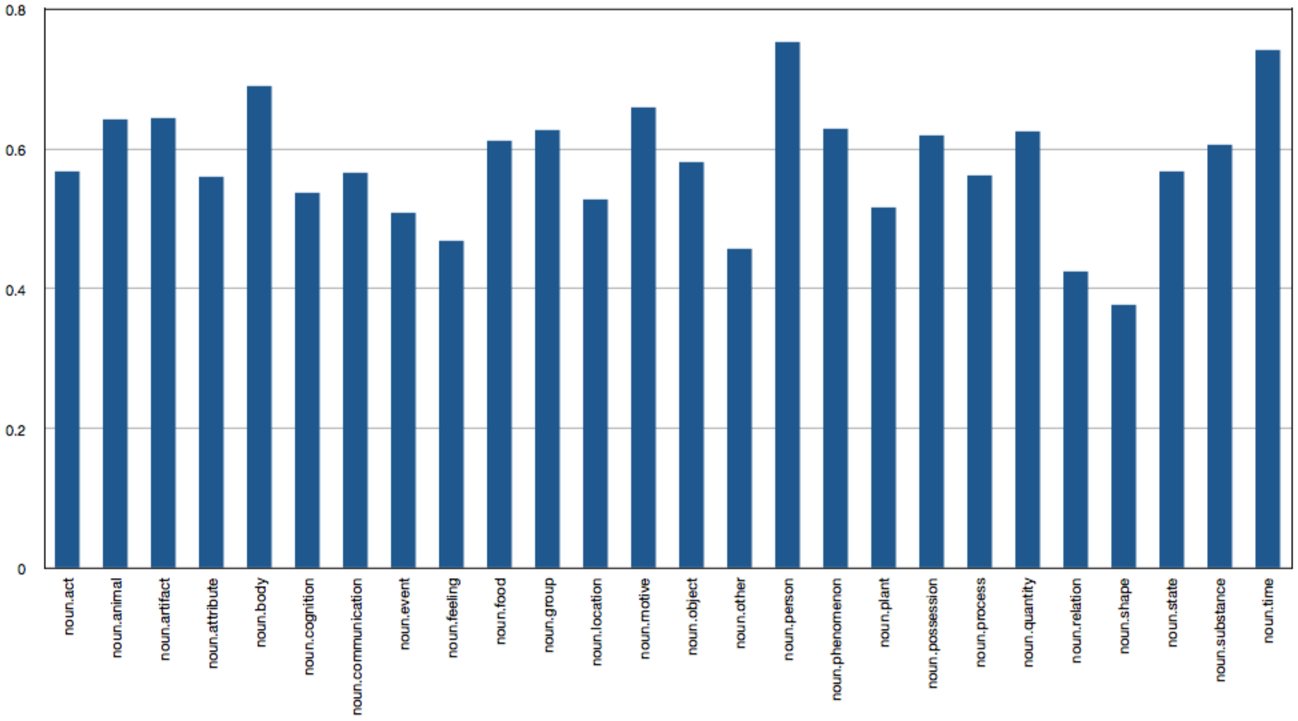

Figure 1: The precision of noun categories