EVALITA 2011

Evaluation of NLP and Speech Tools for Italian

EVALITA 2011

Description and Results of the SuperSense Tagging Task

S. Dei Rossi, G. Di Pietro, M. Simi

Dipartimento di Informatica, Università di Pisa

Introduction

- Why Super Sense tagging?
  - Named Entity Recognition (NER)
    - Simple ontologies: person, organization, location …
    - Limited semantic/syntactic coverage
    - High accuracy
  - Word Sense Disambiguation
    - Identifying WordNet senses
    - tens of thousands of specific "word senses"
    - all open class words covered, domain-independent
    - inadeguate performance

SuperSenses

- SuperSenses
  - Introduced by Ciaramita and Altun (2006)
- WordNet SuperSenses
  - Noun and verb synsets mapped to 41 general semantic classes (lexicographic categories)
  - 26 noun categories; 15 verb categories
- Example:

  "Clara Harris$_{person}$, one of the guests$_{person}$ in the box$_{artifact}$, stood up$_{motion}$ and demanded$_{communication}$ water$_{substance}$"

SuperSenses

1. adj.all
2. adj.pert
3. adv.all
4. noun.Tops
5. noun.act
6. noun.animal
7. noun.artifact
8. noun.attribute
9. noun.body
10. noun.cognition
11. noun.communication
12. noun.event
13. noun.feeling
14. noun.food
15. noun.group
16. noun.location
17. noun.motive
18. noun.object
19. noun.person
20. noun.phenomenon
21. noun.plant
22. noun.possession
23. noun.process
24. noun.quantity
25. noun.relation
26. noun.shape
27. noun.state
28. noun.substance
29. noun.time
30. verb.body
31. verb.change
32. verb.cognition
33. verb.communication
34. verb.competition
35. verb.consumption
36. verb.contact
37. verb.creation
38. verb.emotion
39. verb.motion
40. verb.perception
41. verb.possession
42. verb.social
43. verb.stative
44. verb.weather
45. adj.ppl

Preliminary results

- For English (Ciaramita and Altun, 2006)
  - training on SemCor (Senseval-3)
  - discriminative HMM, trained with an average perceptron algorithm
  - average F-Score on 41 categories: 77.18
- For Italian
  - Picca, Gliozzo, Ciaramita (LREC 2008)
    - trained on MultiSemCor (Bentivoglio et al.)
    - average F-Score on 41 categories: 62,90
  - Attardi, et al. (LREC 2010)
    - trained on ISST-SST
    - average F-Score on 45 categories: 79.10

ISST-SST

- MultiSemCor problems
  - Smaller size (64% of English corpus)
  - Incomplete alignment (sense in Eng., no sense in Ita.)
  - PoS coarseness
  - Word by word translation
- New resource: ISST-SST (G. Attardi, S. Dei Rossi, G. Di Pietro, A. Lenci, S. Montemagni, M. Simi – LREC 2010)
  - Italian Syntactic-Semantic Treebank
  - Large about 300.000 tokens
  - All texts extracted from Italian newspapers

Evalita 2011 ISST-SST (v2)

- Training set
  - About 270.000 tokens from ISST-SST
- Test set
  - The remaining part of ISST-SST (about 30.000 tokens)
  - About 20.000 tokens from the Italian Wikipedia
- Improvements for Evalita 2011
  - All SuperSenses manually revised
  - Expression such as “Croce Rossa”, “Fiona May” and “10 dicembre 1975” considered as single entities
- Evaluation on all 45 SuperSenses:
  - Noun, Verbs and also Adjectives and Adverbs

Task organization

- Two subtasks
  - Closed: only the corpus provided for training
  - Open: any external resource in addition to the corpus provided for training
- The evaluation metrics are quite standard:
  - Tagging accuracy
    - The percentage of correctly classified tokens with respect to the total number of tokens
  - F1-measure
    - the weighted harmonic mean of precision and recall

Participants

- Two participants
  - University of Pisa (UNIPI – Simi et al.)
    - Only Closed Subtask
    - Maximum Entropy classifier and dynamic programming algorithm
  - University of Bari (UNIBA – Basile)
    - Both subtasks
    - Support Vector Machines classifiers and a semantic WordSpace (open subtask only)

Results

Closed Subtask

	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

Open Subtask

	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

Conclusion

- The best performances obtained by the systems of the two teams are very good and very close
  - UNIPI Run 3 – F1: 78.36 vs. UNIBA SVMCat – F1: 76.29
  - F1 and accuracy close to the previous work on Italian SuperSense Tagging (F1 79.10)
- Models learned on the ISST-SST training cope effectively with a different domain (Wikipedia)
  - The performances on the two subparts of the test set are very close
    - UNIPI systems: difference of about 1 point in F1
    - UNIBA systems: difference of about 2-3 points in F1