Overview of the EVALITA 2009 PoS Tagging Task

G. Attardi, M. Simi

Dipartimento di Informatica, Università di Pisa

+ team of project SemaWiki

Outline

- Introduction to the PoS Tagging Task
- Task definition
  - Data sets
  - Evaluation metrics
- Results
  - systems' results
  - discussion
- Conclusion

Introduction

- State of the art:
  - Penn TreeBank data set: 36 categories, accuracy above 97%
  - Evalita 2007 data sets: EAGLES compliant (32 categories); DISTRIB (16 categories), accuracy above 98% with external resources and multi-words lists.
- The challenge for EVALITA 2009:
  - A **larger tag set**: 37 tags with morphological variants, 336 morphed tags (234 actually present in corpus).
  - **Domain adaptation**: training data from newspaper articles (La Repubblica), dev and test data from the Italian Wikipedia

Task definition

- PoS Categories
  - TANL tagset, a Eagles compliant tagset (236 different tags)
  - http://medialab.di.unipi.it/wiki/Tanl_POS_Tagset/
- Data sets
  - Training set: from newspaper La Repubblica;108.874 word forms divided into 3.719 sentences.
  - Development set: from Wikipedia; 5021 forms; 147 sentences
  - Test set: from Wikipedia; 5066 forms; 147 sentences
  - New words: about 17%
- Data format: one token per line, UTF-8 encoded …

Two subtaks

- Open Task: participants can use external resources
- Closed Task: participants are not allowed to use external resources

Evaluation metrics

- *Tagging Accuracy* (TA): the percentage of correctly tagged tokens with respect to the total number of tokens in the Test Set.
- *Unknown Words Tagging Accuracy* (UWTA): tagging accuracy restricted to the unknown words, i.e. tokens present in the Test Set but not in the Training Set.
- TA and UWTA measured w.r.t.:
  - POS: the complete tagset
  - CPOS: fine-grained tags without morphology

Participating groups

Research Team	*Main investigator*	*Affiliation*
SemaWiki	G. Attardi	Dip. di Informatica, Univ. di Pisa, Italy
CST_S¿gaard	A. S¿gaard	Centre for Lang. Tech., Univ. of Copenhagen, Denmark
Gesmundo	A. Gesmundo	Universitˆ di Genova, Italy
Felice-ILC	F. DellÕOrletta	ILC-CNR, Pisa, Italy
Lesmo	L. Lesmo	Dip. di Informatica, Univ. di Turin, Italy
Pianta	E. Pianta	Found. B. Kessler ĞIRST, Trento, Italy
Rigutini	L. Rigutini	Dip. di Ing. Informatica, Univ. di Siena, Italy
Tamburini	F. Tamburini	DSLO, Universitˆ di Bologna, Italy

Open task results

Team	*POS TA*	*CPOS TA*	*POS UWTA*	*CPOS UWTA*	*Rank*
SemaWiki 2	**96.75%**	**97.03%**	**94.62%**	**95.30%**	**1**
SemaWiki 1	96.44%	96.73%	94.27%	95.07%	2
SemaWiki 4	96.38%	96.67%	93.13%	93.81%	3
SemaWiki 3	96.14%	96.42%	92.55%	93.24%	4
Pianta	96.06%	96.36%	92.21%	93.24%	5
Lesmo	95.95%	96.26%	92.33%	93.01%	6
Tamburini 1	95.93%	96.40%	90.95%	92.67%	7
Tamburini 2	95.63%	96.16%	91.07%	92.78%	8

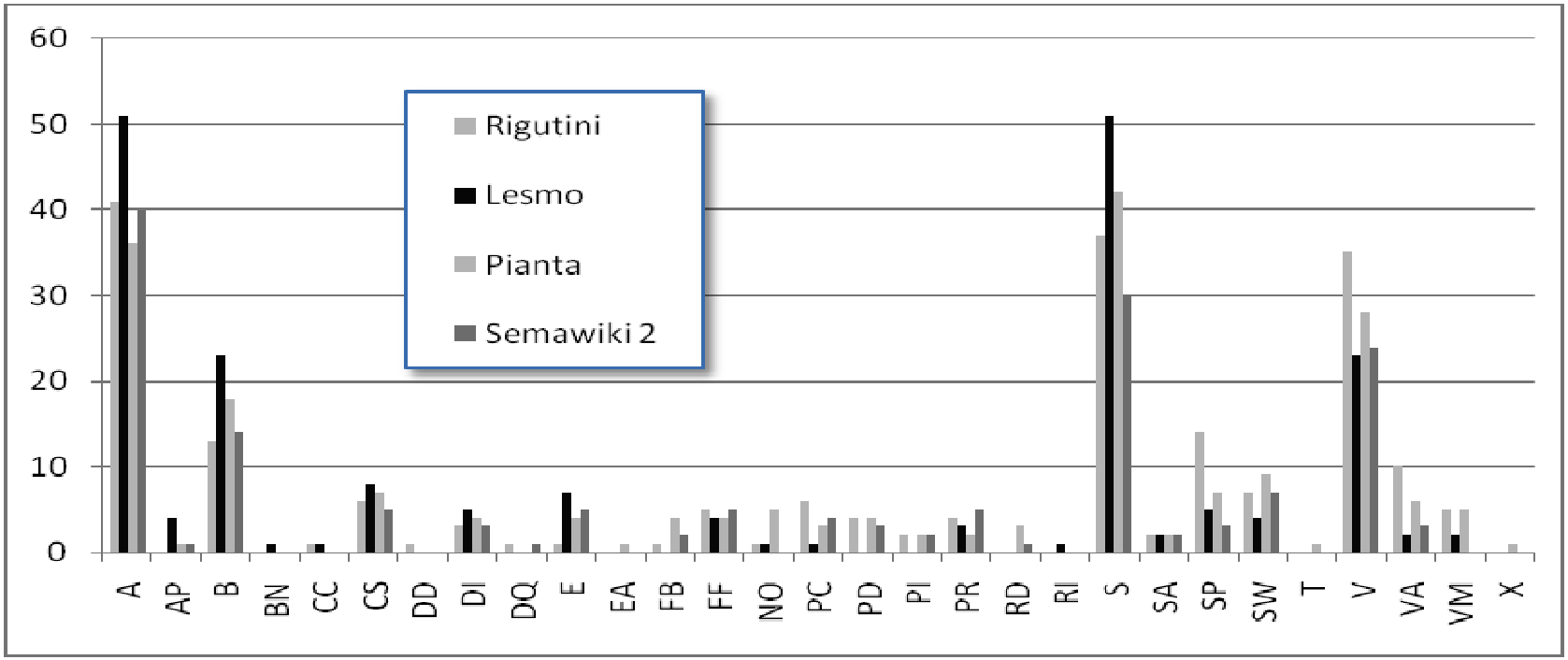
Closed task results

Team	*POS TA*	*CPOS TA*	*POS UWTA*	*CPOS UWTA*	*Rank*
Felice_ILC	**96,34%**	**96,91%**	91,07%	93,36%	**1**
Gesmundo	95,85%	96,48%	**91,41%**	**93,81%**	2
SemaWiki 2	95,73%	96,52%	90,15%	93,47%	3
SemaWiki 1	95,24%	96,00%	87,40%	90,72%	4
Pianta	93,54%	94,10%	85,45%	87,74%	5
Rigutini 2	93,37%	94,15%	86,03%	88,43%	6
Rigutini 3	93,31%	94,15%	86,03%	88,55%	7
Rigutini 4	93,29%	94,17%	85,34%	88.09%	8
Rigutini 1	93,10%	93,76%	84,54%	87,06%	9
CSTS¿gaard 1	91,90%	93,21%	86,03%	89,58%	10
CSTS¿gaard 2	91,64%	93,21%	86,14%	89,92%	11

Summary of approaches

Team	*Type*	*Components*	*Model order*	*n-gram*	*Train. Feat.*	*Search*
SemaWiki	Combination or cascade + rules	Hunpos, TreeTagger	Second	3-gram		Viterbi
CST_S¿gaard	Combination	Brill, TreeTagger, MaxEntropy + combination classifier		1-gram in classifier		none
Gesmundo	Single	Perceptron	First	2-gram bidirectional	515 k	beam-search
Felice-ILC	Combination	HMM, SVM, MaxEntropy	Second	4-gram	SVM: 91400, ME: 939000	Viterbi
Lesmo	Rule based	573 rules				
Pianta	Cascade of 4 classifiers	SVM	First	varying		Viterbi
Rigutini						
Tamburini	Combination	HMM, TBL		2-gram, 3-gram		Viterbi; Brill algorithm

Error Analysis



Conclusions

- We can deal with the higher complexity of the task
  - build practical taggers performing tagging and morphology at the same time
  - reusable across domains
- Comparison with Hunpos, open source implementation of TnT
  - Open task: 96.27% TA, 4th scoring system (same lexicon as 1st)
  - Closed task: 95.14% TA, last scoring system
  - Efficiency: ~0.03min for training, ~0.07min for tagging the Test Set