EVALITA 2009: Ensemble system for Part-of-Speech tagging

Felice Dell'Orletta

Istituto di Linguistica Computazionale - CNR Pisa
felice.dellorletta@ilc.cnr.it

december 12, 2009

General System Description

- **FDO-POS-Tagger** is a combination of six component taggers, with three different algorithms, each of which is used to develop a left-to-right (LR) tagger and a right-to-left (RL) tagger.

- The tagger combine the outputs of the component taggers using one of the following methods:
  - Simple Voting scheme;
  - Machine-learning classifier to identify the correct output among the outputs of the component taggers;
  - Machine-learning classifier to identify the correct POS tag using the outputs of component taggers as features.

Component Taggers

- The first POS tagging algorithm is a popular algorithm for tagging based on Trigrams'n'Tags (TnT) which has readily available open source reimplementation called *Hunpos*. The TnT tagger (Brants, 2000) is an implementation of the Viterbi algorithm for second order Markov model.

- The other two tagging algorithms are based on *ILC-UniPi-tagger* (Dell'Orletta et al. Evalita07). We developed a modular python implementation of this tagger that can use several learning algorithms and provides a simple definition of feature models. In the context of the Evalita 2009 POS tagging task we used Support Vector Machines (SVM) and Maximum Entropy (ME) as learning algorithms.

Constituent Taggers

- These tables show the feature models for the SVM and ME taggers. Only for ME-based taggers we use bigram and trigram features. Right-to-left and left-to-right taggers use the same set of features.

Feature	token
FORM	-2 -1 0 1
FORM_LENGTH	0
FORM_FORMAT	0
FORM_PREFIX	0
FORM_SUFFIX	0
FORM_SHAPE	0
POS	-1

Table: SVM and ME (RL and LR) Felice-ILC-POS-Tagger: feature models.

BIGRAM	$(P_{-1}W_0)$ $(W_{-1}W_0)$ (W_0W_1) (W_1W_2)
TRIGRAM	$(P_{-2}P_{-1}W_0)$ $(W_{-1}W_0W_1)$ $(W_{-2}W_{-1}W_0)$ $(W_0W_1W_2)$

Table: ME (RL and LR) Felice-ILC-POS-Tagger: bigram and trigram features. (W_x=form of token x ; P_x=POS of token x)

Accuracy

- The relative accuracies of the six component taggers:

	H-LR	H-RL	SVM-RL	SVM-LR	ME-RL	ME-LR
development set	92.82	92.72	91.39	91.16	91.19	90.84
test set	95.97	95.55	94.76	94.29	94.25	93.76

Table: Accuracy of the component POS taggers

Complementarity, Disagreement and Additivity rates

- Now we show a series of evaluation measures, proposed by Brill and Wu 1998, to calculate how different the errors of the taggers are.
- We show that the errors the different taggers make are complementary.
- It's clear that if all the taggers made the same errors or if the lower accuracy tagger errors contain all the higher accuracy tagger errors, the tagger would have not improved accuracy through classifier combination.

Complementarity

- Brill and Wu define the *complementary rate* of taggers A and B as:

$$Comp(A,B) = (1 - \frac{\#of\ common\ errors}{\#of\ errors\ in\ A\ only}) * 100$$

- $Comp(A,B)$ measures the percentage of time when tagger A is wrong and that tagger B is correct.

	H-LR	H-RL	SVM-RL	SVM-LR	ME-RL	ME-LR
H-LR	0	15.66	33.84	36.36	34.85	36.36
H-RL	23.74	0	34.70	35.16	37.90	39.27
SVM-RL	49.22	44.57	0	28.29	26.74	32.95
SVM-LR	55.16	49.47	34.16	0	40.93	29.18
ME-RL	54.42	51.94	33.22	41.34	0	31.80
ME-LR	58.96	56.68	43.65	35.18	37.13	0

Table: Complementarity rate. Comp(A,B). Row=A, Column=B

- When the Hunpos left-to-right (H-LR) tagger is wrong, the worst tagger (ME-LR) is correct 36.36% of the time.

- Left-to-right and right-to-left taggers are quite complementary.

Disagreement

- The *Disagree score* for a component tagger A measures the percentage of time that tagger A disagrees with at least one of the other taggers and A is wrong.

	H-LR	H-RL	SVM-RL	SVM-LR	ME-RL	ME-LR
Overall Error Rate	4.03	4.45	5.24	5.71	5.75	6.24
Error Rate When Disagreement	29.70	34.02	42.06	46.80	47.21	52.16

Table: Disagreement rate

- Quoting Brill and Wu:

> *A tagger is much more likely to have misclassified the tag for a word in instances where there is disagreement with at least one of the other classifiers than in the case where all classifiers agree.*

It is interesting to note that when the best tagger (H-LR) disagrees with the others the Hunpos-LR error rate is 29.70%, instead of the overall error rate 4.03%.

Additivity

- This table shows that the tagger complementarity is *additive*.

	H-LR	+H-RL	+SVM-RL	+SVM-LR	+ME-RL	+ME-LR
% of time all are wrong	4.03	3.39	2.38	2.11	1.85	1.74
% Oracle Improvement		15.66	40.91	47.47	54.04	56.57

Table: Additivity rate

- The first row in the table is the additive error rate of an oracle that can choose among all of the possible outputs of component taggers. The second row is the additive oracle improvement.

- If the oracle uses all the six taggers the additive error rate is 1.74 %, (which means) a decrease of 56.57% with respect to the best tagger (4.03%).

Taggers Combination

- These analyses show that it may be possible to obtain an improvement of the accuracy in POS tagging when combining the six component taggers.

- Experiments conducted on Evalita-2009 development set showed that using the machine-learning classifiers methods we do not achieve an improvement in accuracy score compared to the best single tagger, or very slight improvements are obtained. Both SVM and ME machine-learning algorithms have been used for the combination experiments and the training set was created using a ten-fold method: the original training set was splitted into ten parts and for each part we have trained the component taggers on the other parts and then we tagged the excluded one.

- We achieve the best accuracy score using the simple voting scheme method.

Simple Voting Scheme

- This method consists in combining the outputs of the six individual taggers, choosing for each token the part-of-speech that is selected from the largest numbers of taggers. In case of ties between two or more part-of-speeches we choose the one predicted from the best individual model.

	development set	test set
H-LR	92.82	95.97
Voting Combination	93.24	96,34
% error rate reduction	6.27	9.09

Table: Accuracy scores for development and official test sets

- The Simple Voting allows us to obtain an improvement of 0.42% on the development set and 0,37% on the test set. That is, respectively, a relative error rate reduction of 6.27% and 9.09%, relative to the accuracy of the best single tagger (H-LR).

Conclusion

- We have report our participation to the EVALITA 2009 Part-of-Speech Closed task. Our tagger achieved the best score of the competition.
- In this work, most of the time was spent designing and developing the software, which limited the time alloted for optimizing learning algorithm parameters and for selecting the best set of feature models. For this reason, future works should be dedicated to the selection of new feature models in order to improve the accuracy scores of single component taggers and final ensemble systems. Moreover, further methods of combination should be studied.