
Spotlight
on the
Amount of Training Data
in the SRL shared task

Nancy McCracken
Center for Natural Language Processing
Syracuse University
June 30, 2005

Learning curves at CoNLL-2005

- Training data is Propbank, release 1.0
 - Sections 02-21 WSJ of Penn Treebank
- Three groups carried out learning curve experiments on varying amounts of training data
 - Systems are not comparable on other aspects
 - Different features and techniques
- Collected information for each team on the amount of training data for the final system and the amount of time it took to train
 - Thanks to teams who sent additional data
 - Apologies to any teams whose data was missed

Pradhan, Hacıoglu, Krugler, Ward, Martin, Jurafsky system

- Pradhan, S., Hacıoglu, K., Krugler, V., Ward, W., Martin, J., Jurafsky, D., "Support Vector Learning for Semantic Argument Classification", To appear in *Machine Learning* journal, Special issue on Speech and Natural Language Processing, 2005.
- Propbank release of Feb 2004

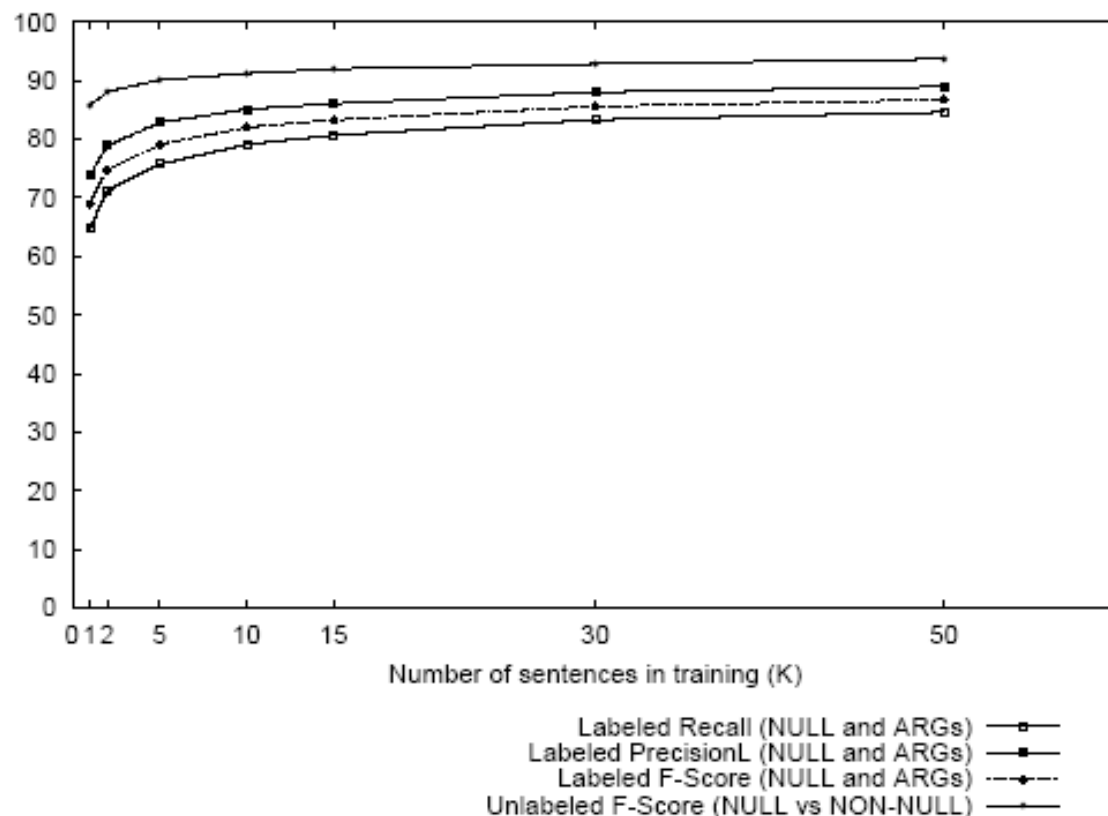
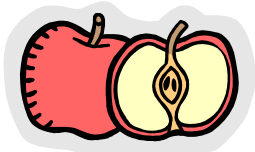
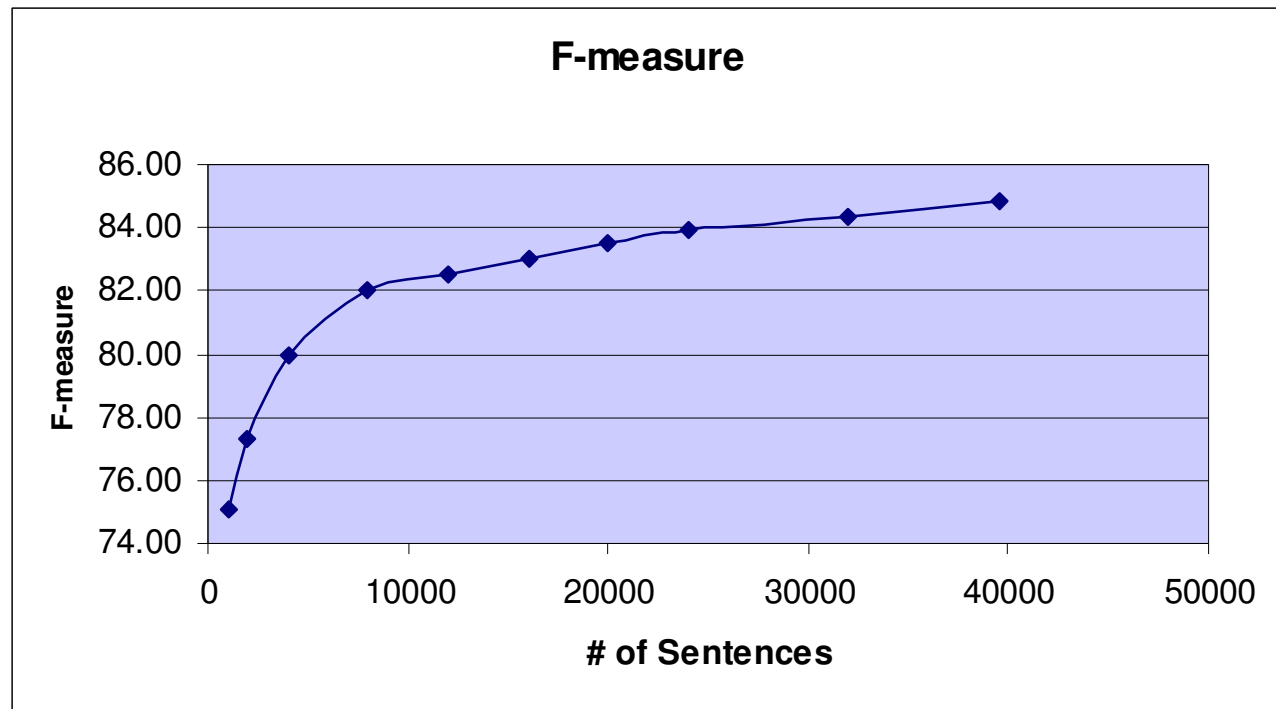


Figure 4. Learning Curve for the task of identifying and classifying arguments using hand-corrected parses.

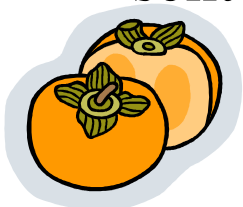


Ozgenicil and McCracken system

- Training the labeling classifier on gold standard parses
- Sentences are in order from 02-21



- Note that F-measure increased approximately 2.5 from 8000 sentences to 32,000 sentences (from 4 to 20 directories)



Che system (Liu, Che, Li, Hu and Liu)

- Combined system trained on sections 02 – 21
- Increments of 4 sections
- Approximately 2,000 sentences in each section

- Note that F-measure increased approximately 3 from 4 to 20 directories

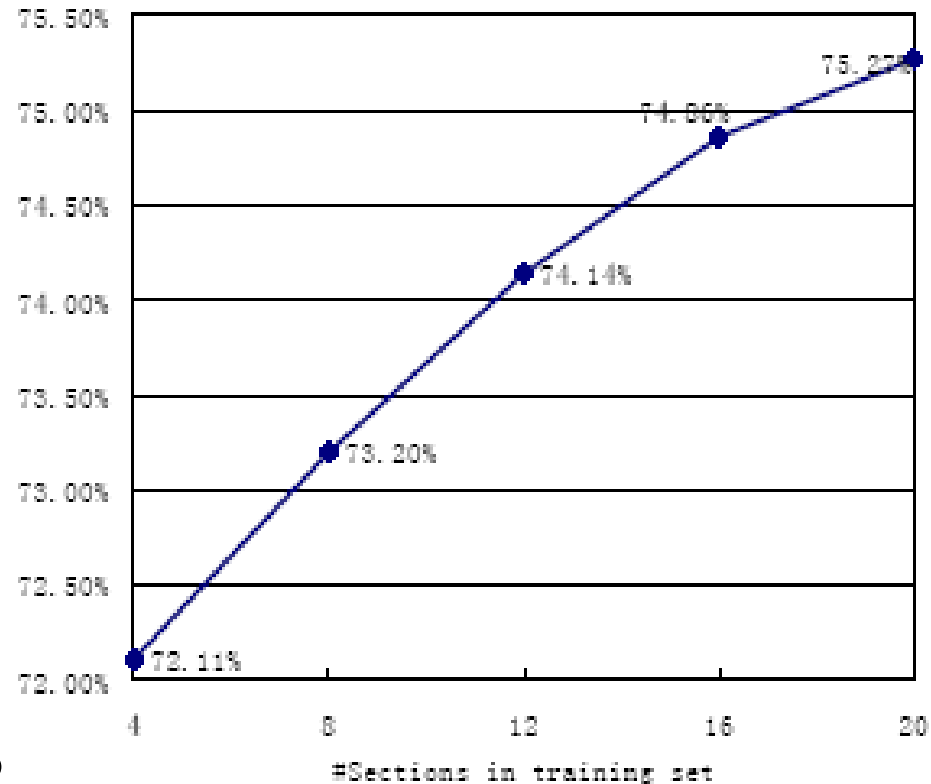
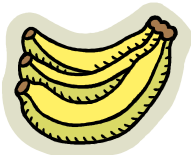


Figure 1: Our SRL system performance curve (of $F_{\beta=1}$) effecting of the training set size.



Training sizes and times

- Most groups used all 20 directories for training.
 - Some groups affected by the amount of training time needed to train on the entire 20 directories
 - Used fewer directories for training
 - Modified techniques
- No attempt made to normalize the training time with respect to the size of machine in the chart on the next page
 - Not enough data
 - Machine sizes ranged from “standard PC” to an 20 processor cluster

Training sizes and times

punayakanok		16.5 hours
haghighi	02-21	9 hrs. 40 mins.
marquez	02-21	about 2 days
tsai	02-21	
che	02-21	20 hours
moschitti	02-08 identifier, 02-21 labeler	about 2.5 days
yi	02-21	
ozgencil	02-11 identifier, 02-21 labeler	30 hours
johansson	02-21 pruning, 15-18 labeler	
cohn	02-21	15 hours
park	02-21	
mitsumori	15-18	
ponzetto	02-21	
lin	02-21	
sutton	02-21	

