

# Evaluation as resource producer

A paradigm that produces high quality language resources

Patrick Paroubek

LIMSI-CNRS  
Dépt. CHM - Groupe LIR  
Bât. 508 Université Paris XI, 91403 Orsay Cedex  
pap@limsi.fr



February 11<sup>th</sup> 2009 / Flarenet / Vienna

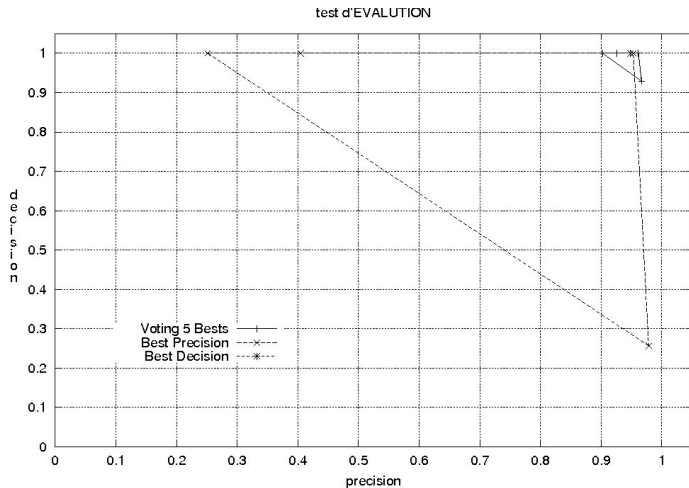
Performance improvement by combining the outputs of many systems using a ROVER (Reduced Output Voting Error Reduction) algorithm.

The acronym and the first experiment of the kind are due to J. Fiscus [1997] in a DARPA/NIST evaluation campaign on speech recognition.

He found out that by aligning the output of the participating speech transcription systems with a dynamic programming algorithm (Allison [1990]) and by selecting the hypothesis which was proposed by the majority of the systems, he obtained better performances than with the best system.

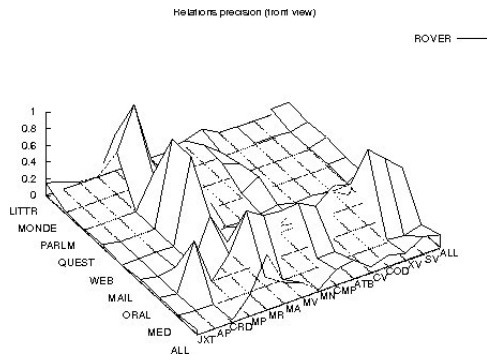
# GRACE-MULTITAG 1996-2000

ROVER best 5 precision-decision improvement over best prec & best dec.

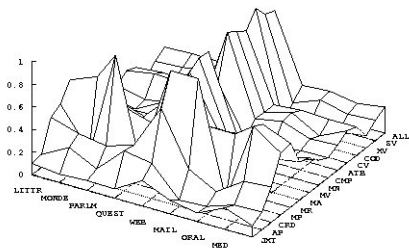


# EASY-EVALDA-TECHNOLANGUE 2002-2005

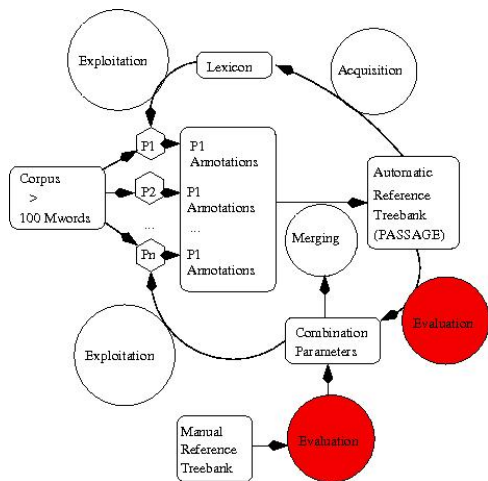
Basic ROVER Relative gain in precision against the best performance / Combining parses to improve quality



difference between recall of union of all participants  
and best recall (P3) at EASY campaign  
(note it is always positive)  
It represent the potential gain in recall for pure combination methods



Evaluation Parametrizes automatic parses combination to produce large sized treebanks



*General Bounds on Statistical Query Learning and PAC Learning with Noise via Hypothesis Boosting*, Javed A. Aslam and Scott E. Decatur (1993)

*“We show that it is possible to improve the accuracy of weak learning algorithms in the Statistical Query model to any arbitrary accuracy...”*

# Conclusion

Evaluation a resource producer

Evaluation : a paradigm that produces high quality validated language resources at relatively low cost.