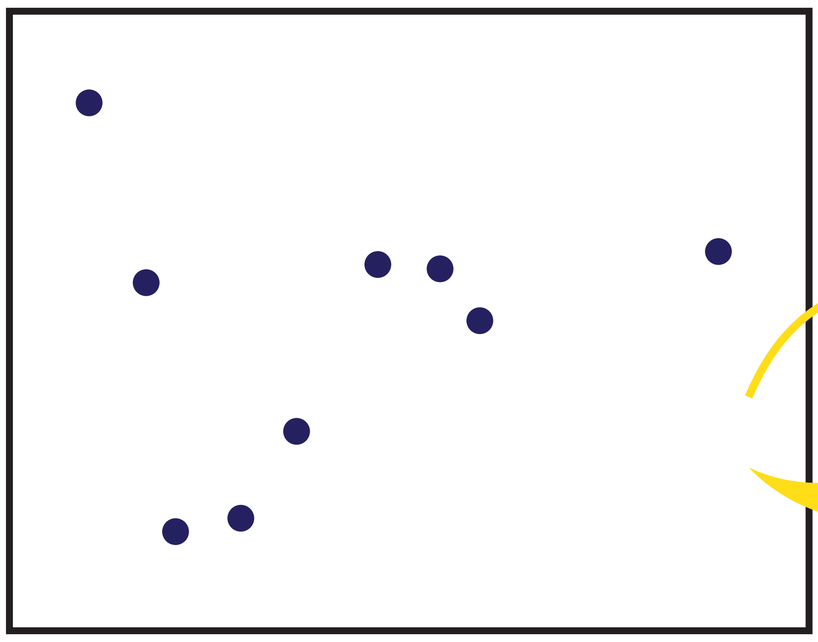
USING DOMAIN SIMILARITY FOR PERFORMANCE ESTIMATION

VINCENT VAN ASCH
WALTER DAELEMANS

CLiPS
UNIVERSITY OF ANTWERP

OBJECTIVE: Trying to find a simple relation between token frequencies in corpora and the cross-domain accuracy of part-of-speech taggers.



Similarity mapping of domains

Imaginative texts ↕ Natural science	88.8
Belief & thought ↕ Imaginative texts	91.6
Social science ↕ Belief & thought	93.1

Accuracy of POS-tagger

Find relation between corpora

Predict accurracy

Training corpus compression

Training corpus selection

Application

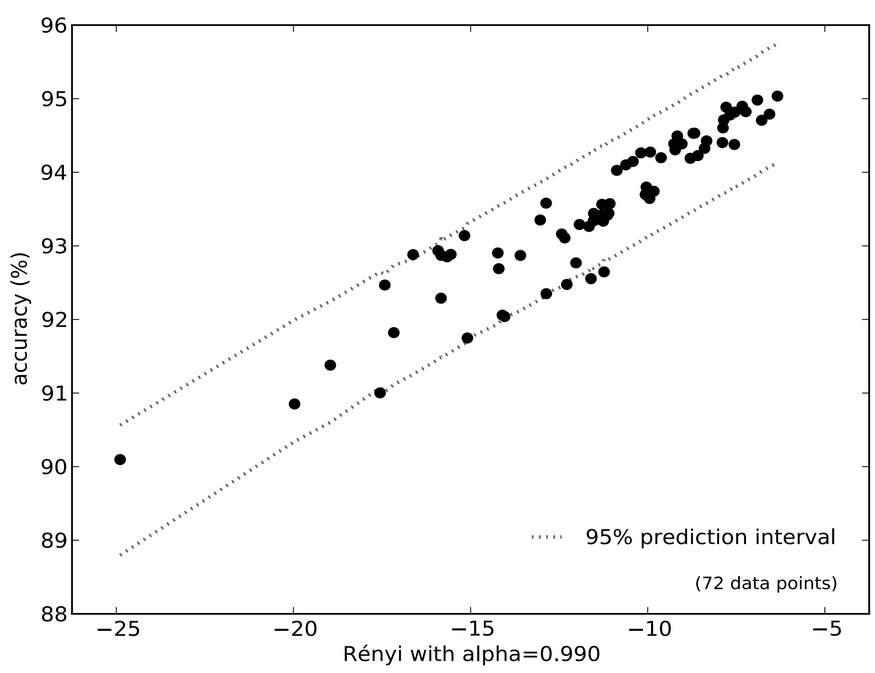
Domains of BNC corpus	Arts	Applied science	Belief & thought	Commerce & finance	Leisure	Imaginative texts	Natural science	Social science	World affairs

Part-of-speech taggers	Rule base - majority	Memory-based - MBT	SVM-based - SVMTool

Experimental design

	Imaginative texts	Natural science
smiled	203	1
sat	260	1
herself	363	1
development	1	161
DNA	1	184
data	1	233

Corpora as token frequencies



$$\frac{1}{\alpha - 1} log_2\left(\sum^{k} p_k^{1-\alpha} q_k^{\alpha}\right)$$

Rényi-divergence with α = 0.99 gave the best linear correlations

Cross-validation experiments indicated:

- Linear relation between accuracy and metric
- Error reduction on accuracy prediction
- Linear relation for in-domain experiments

More information about research at CLiPS is available at www.clips.ua.ac.be

Acknowledgements: This research is financed by the University of Antwerp - GOA and by FWO.