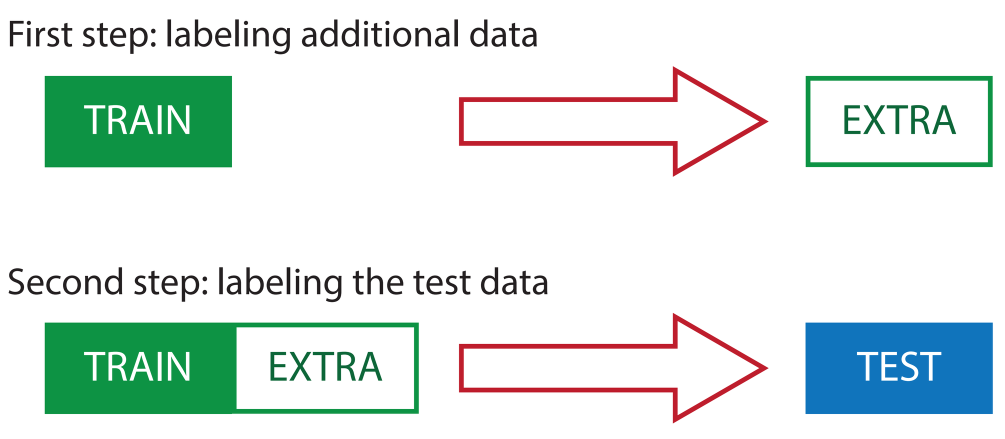
# An analytical approach to similarity measure selection for self-training


Vincent Van Asch and Walter Daelemans

Vincent.VanAsch@ua.ac.be    Walter.Daelemans@ua.ac.be

CLiPS Research Centre, University of Antwerp, Belgium


## 1 Self-training setup

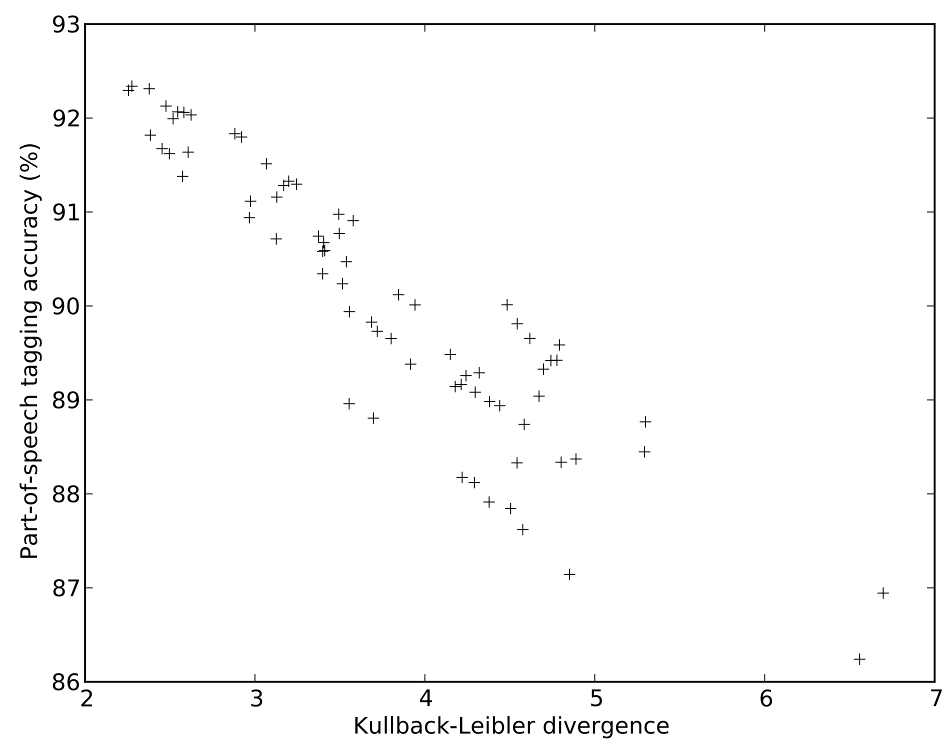



## 2 Research question

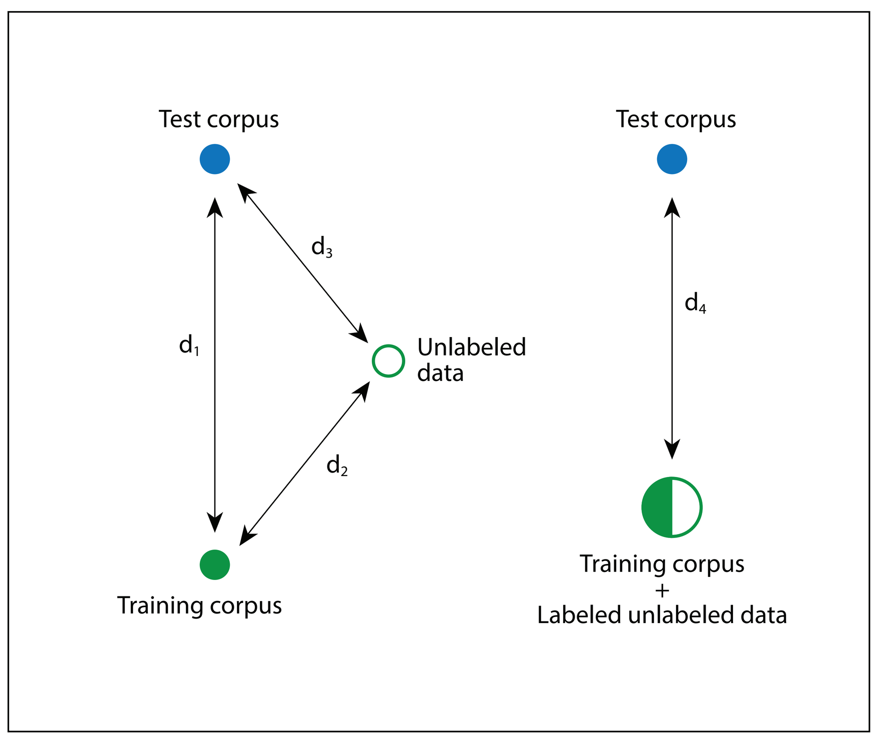
Sagae (2010) argues that self-training is only beneficial if training and test data are sufficiently dissimilar. But how to identify situations for which self-training helps?

## 3 Observations

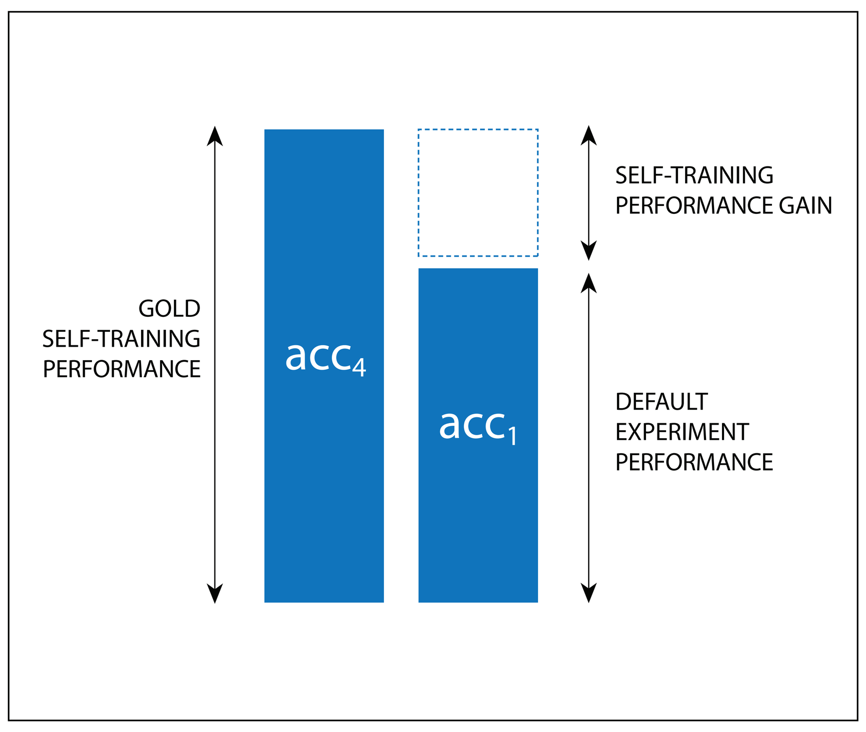
Performance of a part-of-speech experiment is inversely proportional to the dissimilarity of the corpora involved.




The various differences between corpora.




Self-training performance gain expressed as a difference.




## 4 Proposal

Use performance indicator δ based on the similarity score beween test and training corpus ($d_1$) and test corpus and the unlabeled data ($d_3$) to identify good self-training setups:

$$\delta = \frac{\left|\frac{d_1}{d_3} - 1\right|}{\frac{d_1}{d_3} - 1}$$

If δ is +1, gain is expected; if δ is -1 no gain is expected.

## 5 Corpus and machine learner

Part-of-speech tagging experiments using the British National Corpus (2001).

Nine domains, each domain corpus limited to 1,500,000 tokens.

Nine domains, three domains needed per self-training experiment means 504 self-training experiments (74 with performance gain; 430 without).

The machine learner is MBT (Daelemans & van den Bosch, 2005) because of its competitive performance and processing speed.

## 6 Self-training gain prediction

| Name | F-score |
|---|---|
| Rényi | 25.15 - 42.33* |
| Kullback - Leibler | 40.79* |
| Skew | 33.13* - 42.94* |
| sUWR | 38.04* |
| Jensen-Shannon | 41.72* |
| *Baseline* | *25.61* |
| Overlap | 22.09 |

* indicates when performance is significantly (5%) better than baseline. Using approximate randomization testing.

Baseline: assume that all self-training setups lead to performance gain.

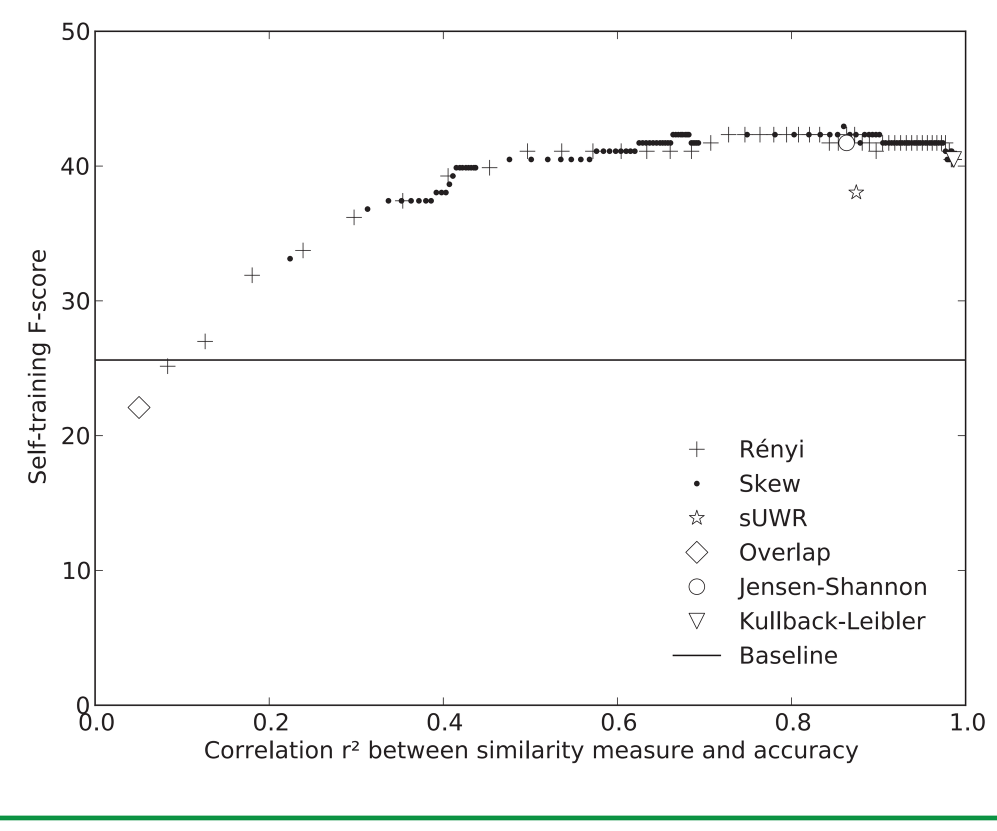
Using the performance indicator δ almost always helps to identify self-training gain.

Rényi and Skew divergence have a parameter α.
Observation: increasing the influence of the test corpus in both similarity measure leads to a decrease in performance.

## 7 Similarity measure selection

| Name | $r^2$ |
|---|---|
| Rényi | 0.083 - 0.987 |
| Kullback - Leibler | 0.986 |
| Skew | 0.224 - 0.985 |
| sUWR | 0.874 |
| Jensen-Shannon | 0.863 |
| Overlap | 0.051 |

Additional research question:
Is the correlation coefficient between similarity and accuracy a good criterion for selecting the best similarity measure?




Observation: a higher correlation coefficient ensures that the performance indicator will be better, but the curve flattens once a certain $r^2$ level is reached.

Most metrics reach the required correlation level, except overlap and some values of the parametrized similarity measures.

## 8 Conclusions

- It is possible to identify self-training setups that will lead to performance gain
- The value of $r^2$ between similarity and accuracy can be used to select a suitable similarity measure
- Weakening the influence of the test corpus increases performance of parametrized similarity measures
- If self-training does not help. The failure may be attributed to the combination of corpora as well as to the possibility that self-training does not help for the task

## Kullback-Leibler

$$\boldsymbol{KL}(P; Q) = \sum_k p_k \log_2\left(\frac{p_k}{q_k}\right)$$

## Rényi divergence

$$\boldsymbol{R}(P; Q; \alpha) = \frac{1}{(\alpha - 1)} \log_2\left(\sum_k p_k^{\alpha} q_k^{1-\alpha}\right)$$

with $\alpha \geq 0$

## Skew divergence

$$\boldsymbol{S}(P; Q; \alpha) = \boldsymbol{KL}\left(Q; \alpha P + (1-\alpha)Q\right)$$

with α in [0, 1]

## Jensen-Shannon

$$\boldsymbol{JS}(P; Q) = \frac{1}{2}\boldsymbol{KL}\left(P; \frac{P+Q}{2}\right) + \frac{1}{2}\boldsymbol{KL}\left(Q; \frac{P+Q}{2}\right)$$

## sUWR

***sUWR***(P; Q) =

the number of tokens that are in test P, but not in train Q divided by the number of tokens in test P

## Overlap

***Overlap***(P; Q) =

the number of tokens that are in train Q, but not in test P divided by the number of tokens in train Q

Universiteit Antwerpen

## Selected references


BNC (2001). The British National Corpus, version 2. Available at www.natcorp.ox.ac.uk

Daelemans & van den Bosch (2005). Memory-based language processing. Studies in Natural Language Processing. Cambridge: Cambridge University Press.

Sagae (2010). Self-training without reranking for parser domain adaptation and its impact on semantic role labeling. Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing. ACL.

Van Asch & Daelemans (2010). Using domain similarity for performance estimation. Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing. ACL.



This research is funded by the Research Foundation Flanders (FWO-project G.0478.10 - Statistical Relational Learning of Natural Language) and made possible through financial support from the University of Antwerp (GOA project BIOGRAPH).


CLiPS


BENELEARN 2013, Nijmegen