Detection target dependent score calibration for language recognition

Raymond W. M. Ng*, Cheung-Chi Leung†, Tan Lee*, Bin Ma† and Haizhou Li†‡

*Department of Electronic Engineering
The Chinese University of Hong Kong

†Institute for Infocomm Research
Singapore

‡ Department of Computer Science and Statistics,
University of Eastern Finland, Finland

ODYSSEY 2010
June 29, 2010

Contents

Contents

LRE task

Given a target language, the task of language recognition is to detect the presence of target in a (testing) trial.

A practical automatic language recognition system (detector) calculates the scores (mostly likelihood) indicating the presence of target, based on which decision is made.

When an erroneous decision is made, a detection cost is incurred. Typical detection cost includes detection misses and false alarms.

Score calibration

Score calibration adjusts the numerical values of scores, which in turn affects detector's decision. The objective is to have a minimum detection cost.

In global calibration, the parameters in the detection cost function, which are specific to an experiment setting, are usually ignored. [Brümmer 2006]

Detection target dependent calibration

Global score calibration:

- transforms the likelihood scores in a global manner
- does not pay special attention to highly confusable trials

In LRE 2009, there are some pairs of related languages.
Detection to these related languages becomes a bottleneck.

- Russian-Ukrainian
- Hindi-Urdu
- Farsi-Dari
- Bosnian-Croatian
- English(American)-English(Indian)

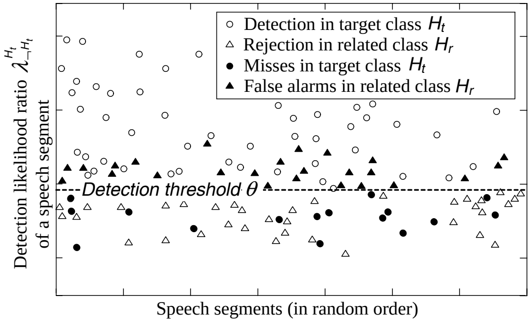
- Will calibration specific to scores of these related language pairs benefit the global cost performance?

Contents

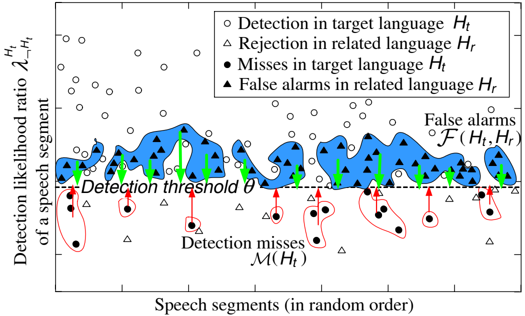
Graphical illustration: Detection based on scores

Testing data in two classes: H_t and H_r
$\lambda^{H_t}_{\neg H_t}$ is the score from the detector, indicating the likelihood H_t

Let k be the index of a test trial, Plot of $\lambda^{H_t}_{\neg H_t}(k)$ against k:



Reduction of total erroneous deviations



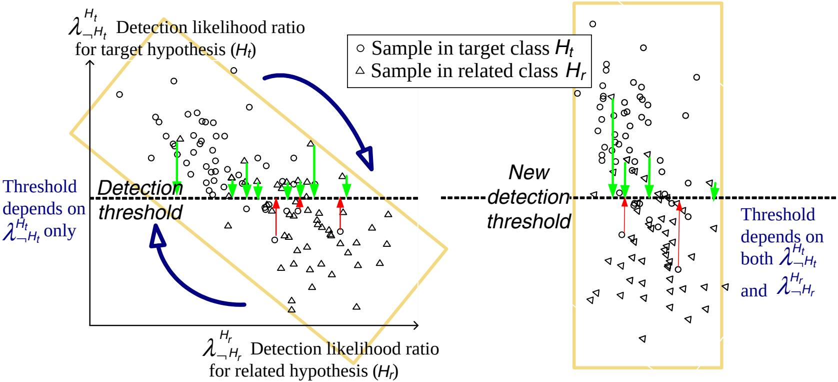
We would like to reduce both sets of detection misses $\mathcal{M}(H_t)$ and false alarms $\mathcal{F}(H_t, H_r)$.

This can be done by minimizing the erroneous deviations, with respect to the detection threshold θ.

Contents

Score adjustment

Hypothesis: Log likelihood ratios for two related languages: $\lambda^{H_t}_{\neg H_t}$ and $\lambda^{H_r}_{\neg H_r}$ contains similar and complementary information.



Total erroneous deviations

Define total erroneous deviations $= \sum_{k=1}^{K} \max\left(y_{H_t}(k)(\lambda_{\neg H_t}^{H_t}(k) - \theta), 0 \right)$

$$y_{H_t}(k) = \begin{cases} 1 & \text{if } k \notin \mathcal{I}(H_t) \\ -1 & \text{if } k \in \mathcal{I}(H_t) \end{cases}$$

- Correct acceptance/rejection: $y_{H_t}(k)(\lambda_{\neg H_t}^{H_t}(k) - \theta) < 0$
- Detection misses: $(\lambda_{\neg H_t}^{H_t}(k) - \theta) < 0; y_{H_t}(k) = -1$
- False alarms: $(\lambda_{\neg H_t}^{H_t}(k) - \theta) > 0; y_{H_t}(k) = 1$

We would like to adjust the detection log likelihood ratio $\lambda_{\neg H_t}^{H_t} \longrightarrow \lambda'^{H_t}_{\neg H_t}$ where the adjusted likelihood could reduce total erroneous deviations

Parameter optimization

Objective function: (with development set) [Boyd 2004]

$$\min_{\alpha_{H_t,H_r}} \sum_{k=1}^{K} \max\left(y_{H_t}(k)(\lambda'^{H_t}_{\neg H_t}(k, \alpha_{H_t,H_r}) - \theta), 0 \right)$$

subject to $|\alpha_{H_t,H_r}| \leq 1$,

$$y_{H_t}(k) = \begin{cases} 1 & \text{if } k \notin \mathcal{I}(H_t) \\ -1 & \text{if } k \in \mathcal{I}(H_t) \end{cases},$$

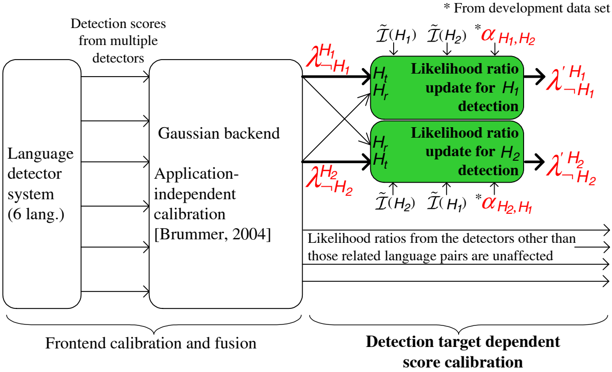
$$\lambda'^{H_t}_{\neg H_t}(k, \alpha_{H_t,H_r}) = \lambda^{H_t}_{\neg H_t}(k) + \alpha_{H_t,H_r} \lambda^{H_r}_{\neg H_r}(k)$$

Evaluation metric: (with evaluation set)

EER of the confusion cost in detecting H_t $[\underset{\theta_{H_t}}{\text{eer}}\, C_{\text{cf}}(H_t)]$, where:

$$C_{\text{cf}}(H_t) = \frac{1}{2} P_{\text{Miss}}(H_t) + \frac{1}{2} P_{\text{FA}}(H_t, H_r)$$

Calibration system setup



Training data: NIST LRE 1996 - 2007 corpora

Evaluation data: NIST LRE 2009 evaluation set(General LR: 10635 trials/23 languages)

Development data: NIST LRE 2007 evaluation set / Excerpts from NIST LRE09 development set (6041 trials/23 languages)

Test duration: 30 seconds

Experimental results with NIST LRE 2009

A relative 5.83% reduction to the EER is achieved

- Bosnian, Croatian confusion cannot be reduced by this method
- In a related language pair, confusion reduction is more significant for the worse performing language

H_t:Target language	H_r:Related language	Original eer $C_{\text{cf}}(H_t)$ θ_{H_t}	Calibrated:2 lang α_{H_t,H_r}	eer $C_{\text{cf}}(H_t)$ θ_{H_t}	
Bosnian	Croatian	30.10%	−0.17	29.82%	
Croatian	Bosnian	31.33%	−0.01	31.05%	
Dari	Farsi	14.87%	−0.49	12.31%	-17% rel.
Farsi	Dari	12.05%	−0.55	11.54%	
Eng(Ame)	Eng(Ind)	16.10%	−0.52	16.04%	
Eng(Ind)	Eng(Ame)	16.38%	−0.74	15.04%	-8% rel.
Hindi	Urdu	28.28%	−0.59	28.77%	
Urdu	Hindi	30.31%	−0.85	29.05%	-4% rel.
Russian	Ukrainian	14.71%	−0.60	10.32%	-30% rel.
Ukrainian	Russian	11.54%	−0.81	9.77%	-15% rel.
Average		**20.57%**		**19.37%**	

Institute for Infocomm Research A*STAR

Contents

Detection to the full set of target languages

Cost function C_{avg} for two target languages:

$$C_{avg} = \frac{1}{2} \sum_{t \in \{1,2\}} \left(p(H_t) P_{miss}(H_t) c_{miss} + \sum_{n \neq t} (1 - p(H_t)) P_{fa}(H_t, H_n) c_{fa} \right)$$

$$c_{miss} = c_{fa} = 1;\ P(H_t) = 0.5$$

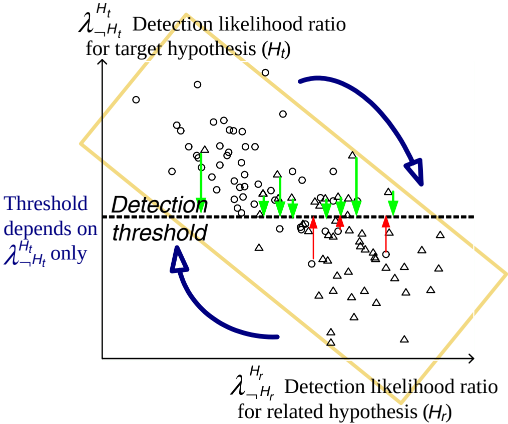
In LRE 2009, there are 23 targets in the general LR task, C_{avg} according to specification:

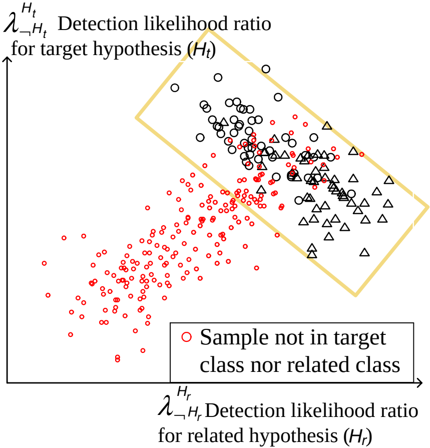
$$C_{avg} = \frac{1}{23} \sum_{t \in \{1...23\}} \left(p(H_t) P_{miss}(H_t) c_{miss} + \sum_{n \in \{1...23\} \setminus t} \frac{1 - p(H_t)}{23 - 1} P_{fa}(H_t, H_n) c_{fa} \right)$$

$$= \frac{1}{23} \sum_{t \in \{1...23\}} C_{detect}(H_t)$$

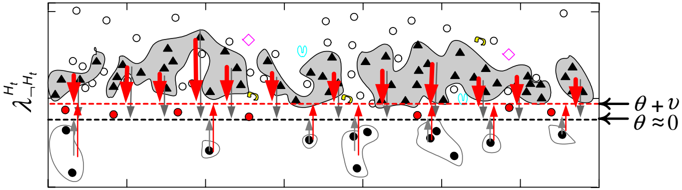
For the detection of each language, there is 1 miss term and 22 false alarm terms to contribute to C_{avg}

Score adjustment with multi-class data





Modification to parameter optimization



- Rule 1: Only select trials which (are likely to) belong to H_t and H_r.
- Rule 2: Weigh the cost of detection misses 22 times heavier
- Rule 3: Shift the reference point for the calculation of total erroneous deviations.

Revised parameter optimization

Revised objective function:

$$\min_{\alpha_{H_t,H_r}} \sum_{k=1}^{K} \max\left(y_{H_t}(k)(\lambda'^{H_t}_{\neg H_t}(k,\alpha_{H_t,H_r}) - (\theta+\upsilon)), 0\right) \longleftarrow \text{Rule 3}$$

$$\text{s.t. } |\alpha_{H_t,H_r}| \leq 1,$$

$$y_{H_t}(k) = \begin{cases} 1 & \text{if } k \notin \mathcal{I}(H_t) \\ -22 & \text{if } k \in \mathcal{I}(H_t) \longleftarrow \text{Rule 2} \end{cases}$$

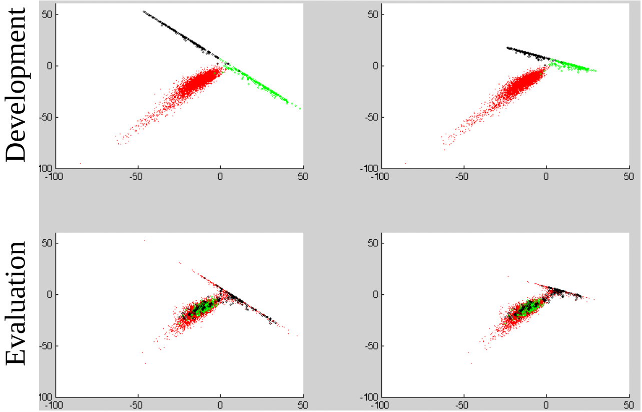
$$\lambda'^{H_t}_{\neg H_t}(k,\alpha_{H_t,H_r}) = \begin{cases} \lambda^{H_t}_{\neg H_t}(k) + \alpha_{H_t,H_r}\lambda^{H_r}_{\neg H_r}(k) & \text{if } k \in \{\tilde{\mathcal{I}}(H_t) \cup \tilde{\mathcal{I}}(H_r)\} \\ \lambda^{H_t}_{\neg H_t}(k) & \text{otherwise} \longleftarrow \text{Rule 1} \end{cases}$$

Evaluation metrics: EER of $C_{avg} =$

$$\frac{1}{23} \sum_{t \in \{1...23\}} \left(p(H_t)P_{miss}(H_t)c_{miss} + \sum_{n \in \{1...23\} \backslash t} \frac{1 - p(H_t)}{23 - 1} P_{fa}(H_t, H_n)c_{fa} \right)$$

Score adjustments for Bosnian detector

Target language: Bosnian; Related language: Croatian; $\alpha = 0.76$

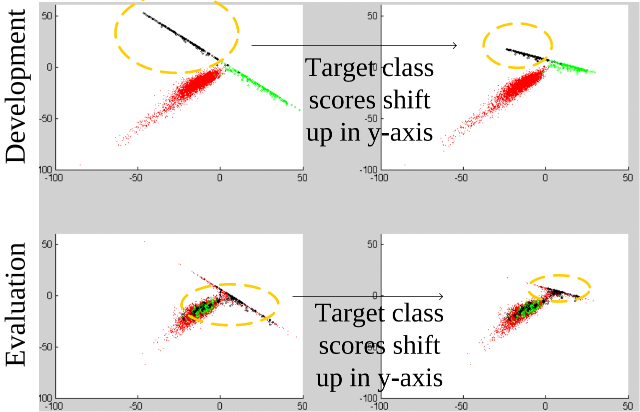


C_{detect} :18.54% (before rotation)

C_{detect} :8.12% (after rotation)

Score adjustments for Bosnian detector

Target language: Bosnian; Related language: Croatian; $\alpha = 0.76$



C_{detect} :18.54% (before rotation)

C_{detect} :8.12% (after rotation)

Experimental results for the full set of target languages

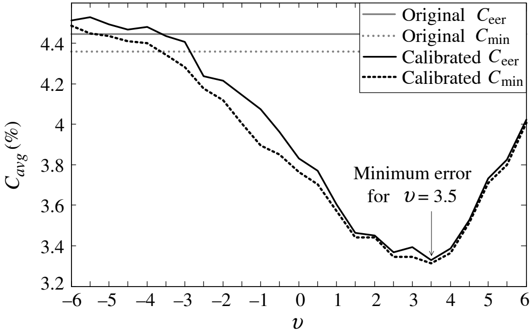
$$C_{avg} = \frac{1}{23} \sum_{t \in \{1...23\}} C_{detect}(H_t)$$

H_t:Target language	H_r:Related language	Original $\underset{\theta}{\text{eer}} C_{\text{detect}}(H_t)$	After calibration α_{H_t,H_r}	After calibration $\underset{\theta}{\text{eer}} C_{\text{detect}}(H_t)$
Bosnian	Croatian	18.54%	0.76	8.12%
Croatian	Bosnian	6.92%	0.43	6.48%
Dari	Farsi	9.07%	0.34	7.03%
Farsi	Dari	3.67%	−0.30	2.65%
Eng(Ame)	Eng(Ind)	4.00%	0.05	3.61%
Eng(Ind)	Eng(Ame)	4.53%	0.13	3.79%
Hindi	Urdu	8.43%	0.62	5.46%
Urdu	Hindi	6.61%	0.67	5.35%
Russian	Ukrainian	5.21%	−0.27	5.35%
Ukrainian	Russian	9.90%	0.76	6.40%
Avg. of 10 “related languages"		7.69%	–	5.42%
Avg. of other 13 languages		1.95%	–	1.72%
Avg. on 23 languages		**4.45%**	–	**3.33%**

Institute for Infocomm Research

A*STAR

Shifting the reference point



When $\upsilon = 3.5$, the lowest C_{avg} is acheived.
Evidence for the detector to prefer fewer detection misses.

Contents

Conclusion

Summary:

- In the language pair detection task for 5 pairs of related languages, a linear combination of the detection scores between the target language and the related language brings 5.8% relative EER reduction
- Revising the parameters for optimization, the application-dependent calibration can be applied to full-set detection. It brings a 25.2% relative EER reduction to 3.33%

Future Work:

- Unsupervised methods to find “related targets”
- Application in other detection tasks

Reference

Selected reference:

[Brümmer 2006] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," in *Computer Speech and Lang.*, vol. 20, no. 2-3, pp. 230-275, 2006.

[Boyd 2004] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

Appendix:Summary of application-independent and dependent calibration

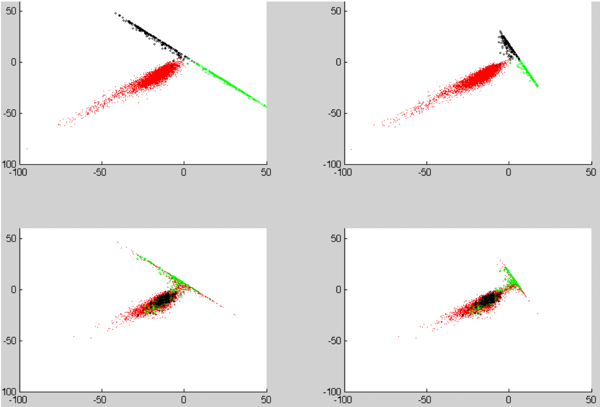
H_t:Target language	$\underset{\theta}{\text{eer}}\ C_{detect}(H_t)$ old method before	after	new method before	after	H_t:Target language	$\underset{\theta}{\text{eer}}\ C_{detect}(H_t)$ old method before	after	new method before	after
Bosnian	6.48%	7.23%	18.54%	8.12%	Cantonese	3.16%	1.35%	1.34%	1.36%
Croatian	5.57%	4.92%	6.92%	6.48%	Mandarin	2.28%	1.45%	1.46%	1.29%
Dari	9.15%	10.20%	9.07%	7.03%	Hausa	2.36%	1.20%	0.91%	0.84%
Farsi	3.37%	2.43%	3.67%	2.65%	Vietnamese	3.48%	2.88%	1.99%	2.02%
Eng(Ame)	3.34%	3.15%	4.00%	3.61%	Portuguese	2.57%	2.04%	1.63%	1.44%
Eng(Ind)	3.90%	5.40%	4.53%	3.79%	Spanish	2.78%	2.78%	3.87%	2.26%
Hindi	8.39%	9.00%	8.43%	5.46%	Amharic	2.74%	1.31%	1.34%	0.89%
Urdu	4.98%	6.79%	6.61%	5.35%	Georgian	4.45%	1.58%	1.55%	1.49%
Russian	3.32%	4.21%	5.21%	5.35%	Korean	1.74%	1.20%	0.96%	0.57%
Ukrainian	6.54%	8.67%	9.90%	6.40%	Pashto	5.92%	5.34%	4.11%	3.46%
Creole	3.58%	2.79%	1.91%	1.81%	Turkish	3.22%	4.09%	1.56%	2.65%
French	5.54%	3.22%	2.74%	2.28%	**Average**	**4.30%**	**4.05%**	**4.45%**	**3.33%**

Appendix:Score adjustments for Croatian detector

Target language: Croatian; Related language: Bosnian; $\alpha = 0.43$

Development

Evaluation



C_{detect} :6.92%
(before rotation)

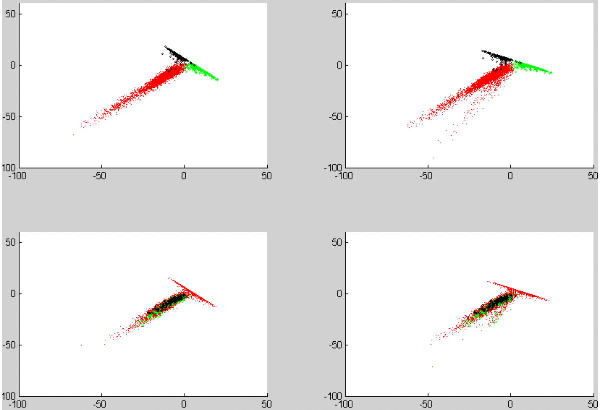
C_{detect} :6.48%
(after rotation)

Appendix:Score adjustments for Dari detector

Target language: Dari; Related language: Farsi; α= 0.34

Development

Evaluation

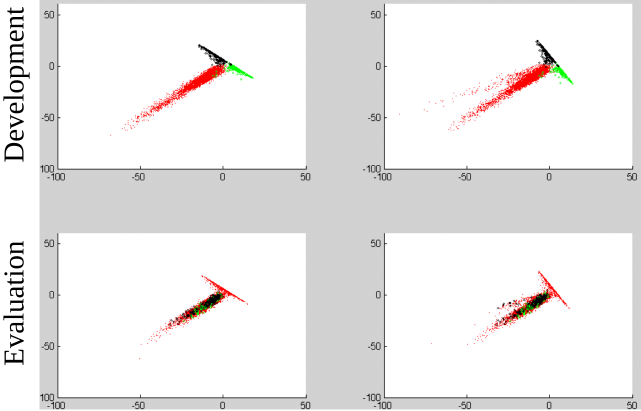


C_{detect} : 9.07% (before rotation)

C_{detect} : 7.03% (after rotation)

Appendix:Score adjustments for Farsi detector

Target language: Farsi; Related language: Dari; $\alpha = -0.30$

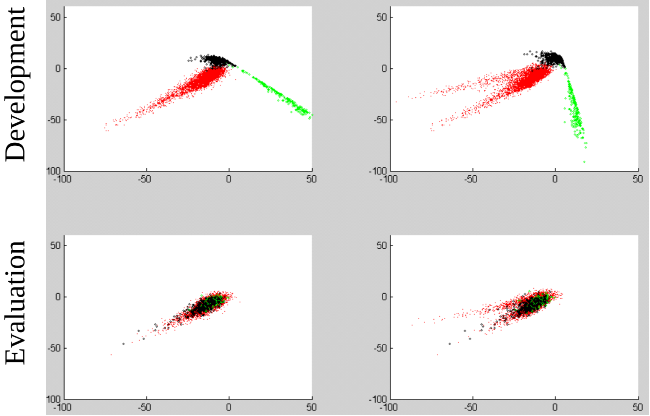


C_{detect} : 3.67%
(before rotation)

C_{detect} : 2.65%
(after rotation)

Appendix:Score adjustments for Russian detector

Target language: Russian; Related language: Ukrainian; $\alpha = -0.27$

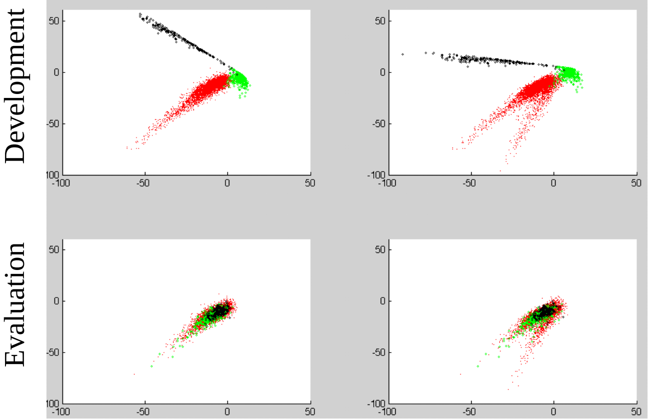


C_{detect} : 5.21%
(before rotation)

C_{detect} : 5.35%
(after rotation)

Appendix:Score adjustments for Ukrainian detector

Target language: Ukrainian; Related language: Russian; $\alpha = 0.76$



C_{detect} : 9.90% (before rotation)

C_{detect} : 6.40% (after rotation)