Can characters reveal your native language? A language-independent approach to native language identification

Radu Tudor Ionescu[1], Marius Popescu[1], Aoife Cahill[2]

[1]Department of Computer Science, University of Bucharest, Bucharest, Romania
[2]Educational Testing Service, Princeton, USA

Introduction

Aim: propose a state of the art method for native language identification (NLI) that works at the character level

Tools:
- kernel-based learning methods
- different string kernels
- multiple kernel learning

Advantages:
- completely language independent: the texts are treated simply as sequences of symbols (strings)
- theory neutral: rather than restricting the feature space according to theoretical or empirical principles, the learning algorithm is left to select the important features
- topic independent: accuracy rate is **30%** higher than the state of the art method in the cross-corpus experiment

String Kernels

- p-spectrum kernel:

$$k_p(s,t) = \sum_{v \in \Sigma^p} \text{num}_v(s) \cdot \text{num}_v(t)$$

- p-grams presence bits kernel:

$$k_p^{0/1}(s,t) = \sum_{v \in \Sigma^p} \text{in}_v(s) \cdot \text{in}_v(t)$$

- p-grams intersection kernel:

$$k_p^{\cap}(s,t) = \sum_{v \in \Sigma^p} \min\{\text{num}_v(s), \text{num}_v(t)\}$$

- The relationship between these kernels:

$$k_p^{0/1}(s,t) \leq k_p^{\cap}(s,t) \leq k_p(s,t)$$

- Normalized:

$$\hat{k}_p(s,t) = \frac{k_p(s,t)}{\sqrt{k_p(s,s) \cdot k_p(t,t)}},$$

$$\hat{k}_p^{0/1}(s,t) = \frac{k_p^{0/1}(s,t)}{\sqrt{k_p^{0/1}(s,s) \cdot k_p^{0/1}(t,t)}},$$

$$\hat{k}_p^{\cap}(s,t) = \frac{k_p^{\cap}(s,t)}{\sqrt{k_p^{\cap}(s,s) \cdot k_p^{\cap}(t,t)}}.$$

Kernel based on Local Rank Distance

- Local Rank Distance (LRD):

$$\Delta_{LRD}(x,y) = \Delta_{left}(x,y) + \Delta_{left}(y,x)$$

$$\Delta_{left}(x,y) = \sum_{i=1}^{|x|-k+1} \min\{|i-j| \text{ such that } 1 \leq j \leq |y| - k + 1 \text{ and } x[i:i+k] = y[j:j+k]\} \cup \{m\}$$

- Kernel:

$$\hat{k}_p^{LRD}(s,t) = e^{-\frac{\Delta_{LRD}(s,t)}{2\sigma^2}}$$

Learning Methods

- Combine kernels by summing kernels and by kernel alignment
- Kernel Ridge Regression (KRR)
- Kernel Discriminant Analysis (KDA)

Experiments

Method	Develop	**10**-fold CV	Test
Ensemble model [Tetreault et al, COLING 2012]	-	**80.9%**	-
KRR and string kernels [Popescu et al, BEA8 2013]	-	**82.6%**	**82.7%**
SVM and word features [Jarvis et al, BEA8 2013]	-	84.5%	**83.6%**
KRR and $\hat{k}_{5-8}^{0/1}$	**85.4%**	**82.5%**	**82.0%**
KRR and $\hat{k}_{5-8}^{\cap}$	**84.9%**	**82.2%**	**82.6%**
KRR and $\hat{k}_{5-8}^{LRD}$	**78.7%**	**77.1%**	**77.5%**
KRR and $\hat{k}_{5-8}^{0/1} + \hat{k}_{5-8}^{LRD}$	**85.7%**	**82.6%**	**82.7%**
KRR and $\hat{k}_{5-8}^{\cap} + \hat{k}_{5-8}^{LRD}$	**84.9%**	**82.2%**	**82.0%**
KRR and $\hat{k}_{5-8}^{0/1} + \hat{k}_{5-8}^{\cap}$	**85.5%**	**82.6%**	**82.5%**
KRR and $a_1\hat{k}_{5-8}^{0/1} + a_2\hat{k}_{5-8}^{\cap}$	**85.5%**	**82.6%**	**82.5%**
KDA and $\hat{k}_{5-8}^{0/1}$	**86.2%**	**83.6%**	**83.6%**
KDA and $\hat{k}_{5-8}^{\cap}$	**85.2%**	**83.5%**	**84.6%**
KDA and $\hat{k}_{5-8}^{LRD}$	**79.7%**	**78.5%**	**79.2%**
KDA and $\hat{k}_{5-8}^{0/1} + \hat{k}_{5-8}^{LRD}$	87.1%	**84.0%**	**84.7%**
KDA and $\hat{k}_{5-8}^{\cap} + \hat{k}_{5-8}^{LRD}$	**85.8%**	**83.4%**	**83.9%**
KDA and $\hat{k}_{5-8}^{0/1} + \hat{k}_{5-8}^{\cap}$	**86.4%**	**84.1%**	**85.0%**
KDA and $a_1\hat{k}_{5-8}^{0/1} + a_2\hat{k}_{5-8}^{\cap}$	**86.5%**	**84.1%**	85.3%
KDA and $\hat{k}_{5-8}^{0/1} + \hat{k}_{5-8}^{\cap} + \hat{k}_{5-8}^{LRD}$	**87.0%**	**84.1%**	**84.8%**

Table : Accuracy rates on TOEFL11 corpus of various classification systems based on string kernels compared with other state of the art approaches.

Method	**5**-fold CV
Ensemble model [Tetreault et al, COLING 2012]	**90.1%**
KRR and $\hat{k}_{5-8}^{0/1}$	**91.2%**
KRR and $\hat{k}_{5-8}^{\cap}$	**90.5%**
KRR and $\hat{k}_{5-8}^{LRD}$	**81.8%**
KRR and $\hat{k}_{5-8}^{0/1} + \hat{k}_{5-8}^{LRD}$	91.3%
KRR and $\hat{k}_{5-8}^{\cap} + \hat{k}_{5-8}^{LRD}$	**90.1%**
KRR and $\hat{k}_{5-8}^{0/1} + \hat{k}_{5-8}^{\cap}$	**90.9%**
KRR and $\hat{k}_{5-8}^{0/1} + \hat{k}_{5-8}^{\cap} + \hat{k}_{5-8}^{LRD}$	**90.6%**
KDA and $\hat{k}_{5-8}^{0/1}$	**90.5%**
KDA and $\hat{k}_{5-8}^{\cap}$	**90.5%**
KDA and $\hat{k}_{5-8}^{LRD}$	**82.3%**
KDA and $\hat{k}_{5-8}^{0/1} + \hat{k}_{5-8}^{LRD}$	**90.8%**
KDA and $\hat{k}_{5-8}^{\cap} + \hat{k}_{5-8}^{LRD}$	**90.4%**
KDA and $\hat{k}_{5-8}^{0/1} + \hat{k}_{5-8}^{\cap}$	**91.0%**
KDA and $\hat{k}_{5-8}^{0/1} + \hat{k}_{5-8}^{\cap} + \hat{k}_{5-8}^{LRD}$	**90.8%**

Table : Accuracy rates on ICLE corpus of various classification systems based on string kernels compared with a state of the art approach. The accuracy rates are reported for the same **5**-fold CV procedure as in [Tetreault et al, COLING 2012].

Method	Big Test
Ensemble model [Tetreault et al, COLING 2012]	**35.4%**
KRR and $\hat{k}_{5-8}^{0/1}$	**66.7%**
KRR and $\hat{k}_{5-8}^{\cap}$	**67.2%**
KRR and $\hat{k}_{5-8}^{0/1} + \hat{k}_{5-8}^{\cap}$	67.7%
KRR and $a_1\hat{k}_{5-8}^{0/1} + a_2\hat{k}_{5-8}^{\cap}$	67.7%
KDA and $\hat{k}_{5-8}^{0/1}$	**65.6%**
KDA and $\hat{k}_{5-8}^{\cap}$	**65.7%**
KDA and $\hat{k}_{5-8}^{0/1} + \hat{k}_{5-8}^{\cap}$	**66.2%**
KDA and $a_1\hat{k}_{5-8}^{0/1} + a_2\hat{k}_{5-8}^{\cap}$	**66.2%**

Table : Accuracy rates on TOEFL11-Big corpus of various classification systems based on string kernels compared with a state of the art approach. The systems are trained on the TOEFL11 corpus and tested on the TOEFL11-Big corpus.

Conclusion

- Results show state of the art accuracy rates for the NLI task
- Intersection kernel successfully used for the first time in text categorization

Future Work

- Analyze the discriminating features selected by the classifier
- Offer a (better) explanation of why this approach works so well

raducu.ionescu@gmail.com acahill@ets.org popescunmarius@gmail.com