Recent Advances in Natural Language Processing
7-13 September, Hissar, Bulgaria

NLEL_UPV_Autoritas participation at Discrimination between Similar Languages (DSL) 2015 Shared Task

Raül Fabra Boluda, Francisco Rangel and Paolo Rosso

Introduction

The objective is to discriminate among similar languages o regional varieties of a given language. We have approached the task with a probabilistic method in two different ways:

- Two-step system:
  - First, determine the language group.
  - Then, predict the variety locally to each group.
- Single-step system: Multiclass classifier to predict directly the variety.

DSLCC v2.0 Dataset

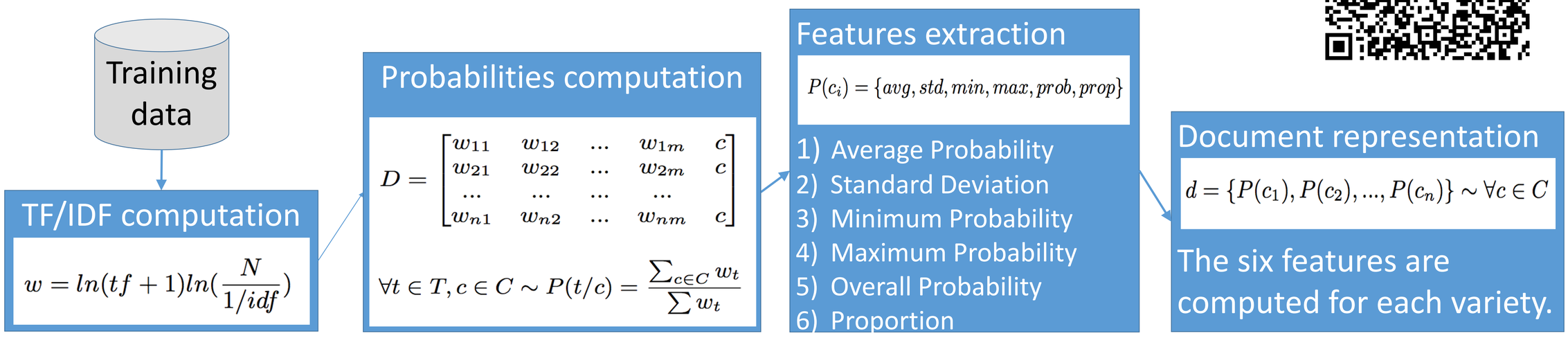
A collection of news in 13 languages/varieties plus a group of "unknown" languages.

Training	Development	Test *
252,000	28,000	14,000

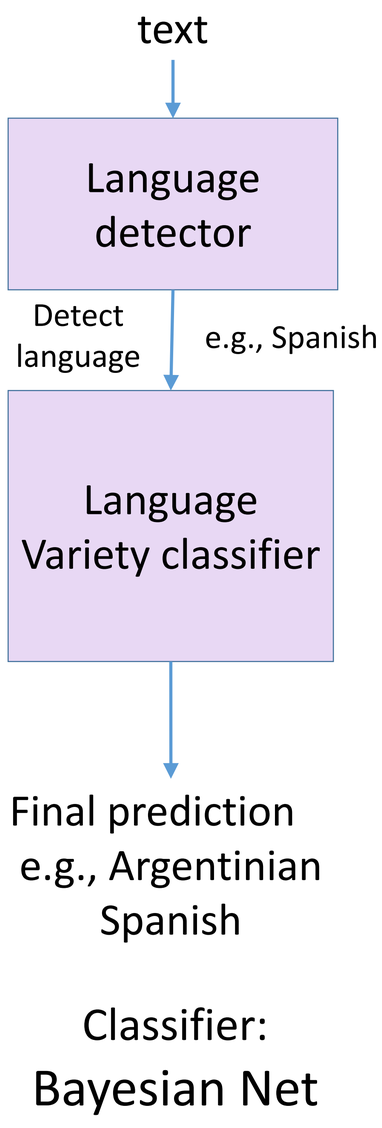
Number of instances per set

* There are two Test sets, A and B. Both are the same, but Test B has no Named Entities.

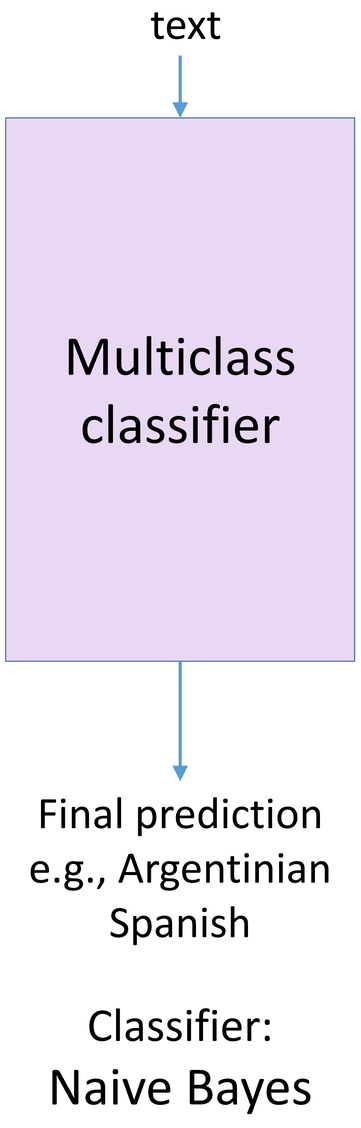
Feature Generation: Probabilistic Method



Two-step System



Single-step System



Experimental Results

Language	Two-Step System			Single-Step System		
	Accuracy			Accuracy		
	Devel.	Test A	Test B	Devel.	Test A	Test B
Bg	99,80	99,90	99,80	98,15	97,50	95,10
Mk	100,00	99,90	100,00	98,95	98,20	98,20
es-ES	88,00	84,70	79,50	87,55	84,80	48,70
es-AR	87,50	88,00	87,70	67,05	70,00	74,10
pt-PT	88,60	87,40	94,00	82,15	81,20	58,30
Pr-BR	90,10	90,03	68,50	72,45	72,50	65,90
Bs	78,35	78,00	74,40	55,70	54,30	86,20
Hr	86,15	85,80	85,40	80,85	78,88	13,10
Sr	86,40	86,40	82,70	74,40	74,70	7,80
Id	99,40	99,40	92,90	97,75	97,60	92,00
My	99,45	99,20	99,50	94,25	93,60	97,60
Cz	99,70	99,80	99,40	98,45	98,40	94,40
Sk	99,60	99,30	99,60	98,80	97,60	79,30
Xx	99,90	99,90	99,70	98,55	98,50	98,80
Overall	93,07	92,71	90,22	86,08	85,57	72,11
Overall*		91,84	89,56		64,04	62,78

* Official results in the competition. They are lower due to a bug in the probabilities computation. Furthermore, some features were considered in wrong order.

Conclusions

- Better results with the two-step system.
- South-Western Slavic languages were the hardest to identify, followed by Spanish and Portuguese.
- Good results for Austraneasian Languages with both approaches.

NLEL Natural Language Engineering Lab
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
autoritas