Comparing Approaches to the Identification of Similar Languages

Marcos Zampieri[1,2], Binyam Gebrekidan Gebre[3], Hernani Costa[4], Josef van Genabith[1,2]

Saarland University[1], DFKI[2]

Max Planck Computing and Data Facility[3], University of Malaga[4]

The Task

- This paper describes the submission made by the MMS team to the Discriminating between Similar Languages (DSL) shared task 2015.
  - ▷ We participated in the closed submission (Test sets A and B).
  - ▷ Used three different systems based.

Related Work

- Malay and Indonesian (Ranaivo-Malancon, 2006)
- Chinese Varieties (Huang and Lee, 2008)
- Portuguese Varieties (Zampieri and Gebre, 2012)
- English Varieties (Lui and Cook, 2013)
- Persian and Dari (Malmasi et al., 2015)

Corpus and Approaches

- We used only the DSLCC v. 2.0. provided by the shared task organizers.

Language/ Variety	Code	Language/ Variety	Code
Bosnian	*bs*	Brazilian Portuguese	*pt-BR*
Croatian	*hr*	European Portuguese	*pt-PT*
Serbian	*sr*	Argentine Spanish	*es-AR*
Indonesian	*id*	Castilian Spanish	*es-ES*
Malay	*my*	Macedonian	*mk*
Czech	*cz*	Bulgarian	*bg*
Slovak	*sk*	Unknown	*xx*

- We approached the task using three different systems:
  - ▷ *Run 1* uses Logistic Regression with TF-IDF Weighting. Tegularisation parameter to 100.0 (Gebre et al., 2013).
  - ▷ *Run 2* uses Support Vector Machine classifier with TF-IDF Weighting (Gebre et al., 2013).
  - ▷ *Run 3* uses a simple, efficient and fast method that combines Laplace smoothing and a probabilistic classifier called likelihood estimation (Zampieri and Gebre, 2012).

Results

Language/Variety	Test Set A	Test Set B
Bosnian	83.5%	76.6%
Croatian	91.8%	92.2%
Serbian	93.9%	90.7%
Indonesian	99.2%	97.5%
Malay	99.4%	99.5%
Czech	100%	99.9%
Slovak	100%	100%
Brazilian Portuguese	93.6%	90.5%
European Portuguese	93.0%	86.7%
Argentine Spanish	91.2%	89.2%
Castilian Spanish	94.8%	94.5%
Macedonian	100%	100%
Bulgarian	100%	100%
Unknown	100%	99.8%

System Performance

Run	Test Set A	Test Set B
Run 1	94.09%	**92.77%**
Run 2	**95.24%**	**92.77%**
Run 3	94.07%	92.47%
Rank	2nd *out of 9*	4th *out of 7*

	bg	bs	cz	es-AR	es-ES	hr	id	mk	my	pt-BR	pt-PT	sk	sr	xx
bg	2000	0	0	0	0	0	0	0	0	0	0	0	0	0
bs	0	1578	0	0	0	241	0	0	0	0	0	0	181	0
cz	0	0	2000	0	0	0	0	0	0	0	0	0	0	0
es-AR	0	0	0	1774	226	0	0	0	0	0	0	0	0	0
es-ES	0	0	0	227	1773	0	0	0	0	0	0	0	0	0
hr	0	132	0	0	0	1841	0	0	0	0	0	0	26	1
id	0	0	0	0	0	0	1979	0	21	0	0	0	0	0
mk	0	0	0	0	0	0	0	2000	0	0	0	0	0	0
my	0	0	0	0	0	0	30	0	1970	0	0	0	0	0
pt-BR	0	0	0	0	0	0	0	0	0	1826	174	0	0	0
pt-PT	0	0	0	0	0	0	0	0	0	222	1778	0	0	0
sk	0	0	0	0	0	0	0	0	0	0	0	2000	0	0
sr	0	86	0	0	0	41	0	0	0	0	0	0	1873	0
xx	0	0	0	0	0	0	0	0	0	0	0	0	0	2000

	bg	bs	cz	es-AR	es-ES	hr	id	mk	my	pt-BR	pt-PT	sk	sr	xx
bg	2000	0	0	0	0	0	0	0	0	0	0	0	0	0
bs	0	1661	0	0	0	193	0	0	0	0	0	0	146	0
cz	0	0	2000	0	0	0	0	0	0	0	0	0	0	0
es-AR	0	0	0	1796	204	0	0	0	0	0	0	0	0	0
es-ES	0	0	0	209	1791	0	0	0	0	0	0	0	0	0
hr	0	135	0	0	0	1843	0	0	0	0	0	0	21	1
id	0	0	0	0	0	0	1988	0	12	0	0	0	0	0
mk	0	0	0	0	0	0	0	2000	0	0	0	0	0	0
my	0	0	0	0	0	0	19	0	1981	0	0	0	0	0
pt-BR	0	0	0	0	0	0	0	0	0	1844	156	0	0	0
pt-PT	0	0	0	0	0	0	0	0	0	166	1834	0	0	0
sk	0	0	1	0	0	0	0	0	0	0	0	1999	0	0
sr	0	86	0	0	0	41	0	0	0	0	0	0	1891	0
xx	0	0	0	0	0	0	0	0	0	0	0	0	0	2000

	bg	bs	cz	es-AR	es-ES	hr	id	mk	my	pt-BR	pt-PT	sk	sr	xx
bg	2000	0	0	0	0	0	0	0	0	0	0	0	0	0
bs	0	1623	0	0	0	198	0	0	0	0	0	0	179	0
cz	0	0	2000	0	0	0	0	0	0	0	0	0	0	0
es-AR	0	0	0	1623	377	0	0	0	0	0	0	0	0	0
es-ES	0	0	0	88	1912	0	0	0	0	0	0	0	0	0
hr	0	205	0	0	0	1746	0	0	0	0	0	0	49	0
id	0	0	0	0	0	0	1980	0	20	0	0	0	0	0
mk	0	0	0	0	0	0	0	2000	0	0	0	0	0	0
my	0	0	0	0	0	0	8	0	1992	0	0	0	0	0
pt-BR	0	0	0	0	0	0	0	0	0	1867	133	0	0	0
pt-PT	0	0	0	0	0	0	0	0	0	236	1764	0	0	0
sk	0	0	0	0	0	0	0	0	0	0	0	2000	0	0
sr	0	107	0	0	0	36	0	0	0	0	0	0	1857	0
xx	5	2	0	5	7	3	0	0	0	0	0	0	2	1976

References

Chu-ren Huang and Lung-hao Lee. 2008. Contrastive approach towards text source classification based on top-bag-of-word similarity. In Proceedings of PACLIC, pages 404–410.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskens. 2013. Improving native language identification with tf-idf weighting. In Proceedings of the 8th BEA workshop, Atlanta, USA.

Marco Lui and Paul Cook. 2013. Classifying English documents by national dialect. In Proceedings of Australasian Language Technology Workshop, pages 5–15.

Shervin Malmasi, Eshrag Refaee, and Mark Dras. 2015. Arabic Dialect Identification using a Parallel Multidialectal Corpus. In Proceedings of PACLING 2015, pages 209–217, Bali, Indonesia, May.

Bali Ranaivo-Malancon. 2006. Automatic identification of close languages - case study: Malay and Indonesian. ECTI Transactions on Computer and Information Technology, 2:126–134

Marcos Zampieri and Binyam Gebrekidan Gebre. 2012. Automatic identification of language varieties: The case of Portuguese. In Proceedings of KONVENS, pages 233–237.

Marcos: marcos.zampieri@uni-saarland.de