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

Raül Fabra-Boluda[1], **Francisco Rangel**[1,2], and **Paolo Rosso**[1]
[1] Natural Language Enginering Lab., Universitat Politècnica de València, Spain
[2] Autoritas Consulting, S.A., Spain
rfabra@dsic.upv.es, prosso@dsic.upv.es,
francisco.rangel@autoritas.es

Abstract

In this paper we describe the participation of the Natural Language Engineering Lab (NLEL) - Universitat Politècnica de València and Autoritas Consulting team in the Discrimination between Similar Languages (DSL) 2015 shared task. We have participated both in open and close submissions. Our system for the open submission performs in two steps. Firstly, we apply a language detector to identify the distinct groups corresponding to families of languages/dialects, and then we distinguish between varieties with a probabilistic method. For the close submission, we implemented our probabilistic method in a multi-class classifier for all the language varieties together. Although our results on the development set were quite promising (93.07% and 86.08% respectively), a software bug (that we have detected only after the submission) dropped considerably our results in the final testing.

1 Introduction

The automatic language identification task aims to determine the language of a given text. The performance on this task is pretty high with long texts (Shuyo, 2010), but it becomes harder when texts are shorter. This may occur in social media scenarios like Twitter (Carter et al., 2013). Furthermore, in social media we may want to go beyond the language scope to identify also dialects or varieties. The objective of the language variety identification is to determine the regional variety of a given language. For example, to know whether a Spanish text is Peninsular, Argentinian, Mexican, and so forth.

Language variety identification may be classified as an author profiling task. Author profiling aims at identifying the linguistic profile of an author on the basis of her writing style. The objective is to determine author's traits such as age, gender, native language, personality traits or language varieties, among others. It is noteworthy the interest in author profiling since 2013, as can be seen in the number of shared tasks: *i)* Age and gender identification at the Author Profiling task at PAN[1] at CLEF 2013 (Rangel et al., 2013) and 2014 (Rangel et al., 2014). In PAN 2015 (Rangel et al., 2015) personality recognition is also treated; *ii)* native language identification at BEA-8 workshop at NAACL-HLT 2013[2] (Tetreault et al., 2013); *iii)* personality recognition at ICWSM 2013[3]; *iv)* Workshop on Language Technology for Closely Related Languages and Language Variants at EMNLP2014[4]; *v)* VarDial Workshop at COLING 2014[5] - Applying NLP Tools to Similar Languages, Varieties and Dialects and *vi)* LT4VarDial - Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialect[6] (Zampieri et al., 2014).

DSL is a hot research topic. The authors in (Sadat et al., 2014) researched the identification of Arabic varieties in blogs and forums. They used character n-grams and Support Vector Machines, and reported accuracies between 70-80% in a 10-fold cross-validation evaluation. Similarly, in (Zampieri and Gebre, 2012) the authors collected 1.000 news articles in two Portuguese varieties: Portugal and Brazil. They used word n-grams and character n-grams and reported accuracies over 90% in a 50-50 split evaluation. They used language probability distributions with log-likelihood function for probability estimation.

[1] http://pan.webis.de
[2] https://sites.google.com/site/nlisharedtask2013/
[3] http://mypersonality.org/wiki/doku.php?id=wcpr13
[4] http://alt.qcri.org/LT4CloseLang/index.html
[5] http://corporavm.uni-koeln.de/vardial/sharedtask.html
[6] http://ttg.uni-saarland.de/lt4vardial2015/dsl.html

In (Maier and Gómez-Rodrıguez, 2014), the authors collected tweets in four different Spanish varieties: Argentina, Colombia, Mexico and Spain. They used four types of features combined with a meta-classifier: character n-gram with frequency profiles, character n-gram language models, LZW compresssion and syllable-based language models. The reported accuracies were between 60-70% in a cross-validation evaluation.

It is also interesting to analyse the submitted systems to the LT4VarDial task. In the system presented in (Goutte et al., 2014) the authors approached the task in two steps. First, it predicted the language group with a 6-way probabilistic classifier. Then, the variety was predicted with a voting combination of discriminative classifiers. They used character and word n-grams and reported 95.71% of accuracy. The system presented in (Porta and Sancho, 2014) used a hierarchical classifier based on maximum-entropy classifiers. The first level predicted the language group and the second the language variety within the predicted group. They experimented with character and word n-grams, together with a list of words which exclusively belong to each language variety. The reported accuracy was 92.6%. The authors in (Purver, 2014) used linear Support Vector Machines with character and word n-grams. They analysed in depth how the cost parameter influenced the classification results, and reported an overall accuracy over 95% after fixing a bug. The system reported in (King et al., 2014) combined character and word n-grams with feature selection techniques such as Information Gain and Parallel Text Feature Extraction. The authors reported that Naive Bayes performed better than Support Vector Machines and Logistic Regression. In (Lui et al., 2014), the authors devoted their research to explore novel methods for DSL. They obtained their best result using their `langid.py` tool (Lui and Baldwin, 2012), with a 91.80% of accuracy.

Our interest in DSL goes beyond the use of features such as n-grams. Our objective is to better understand the linguistic differences between varieties as well as the relationship to other author profiling tasks. In (Franco-Salvador et al., 2015), we approached the DSL task with distributed representations. We also compared with Emograph (Rangel and Rosso, 2015a), a graph-based approach which obtained competitive accuracies with PAN datasets (Rangel and Rosso, 2015b) in the age and gender author profiling tasks. In this paper we describe our participation at the DSL 2015 shared task (Zampieri et al., 2015). We approached the task by proposing a probabilistic method which tries to capture lexical differences between varieties.

2 Identifying Language Varieties

We participated in both open and close tasks. Our objective was to compare the performance of our approach when dividing the identification in two steps against learning all varieties together.

For the open submission we have developed a two-step method. The first step consists in the identification of language groups by means of a language detector. We use the `ldig` language detector developed in (Shuyo, 2010). The author computed character n-grams from Wikipedia abstracts and used Naive Bayes as machine learning algorithm. The reported accuracies are about 99.1% for up to 53 languages.

In the second step, for each language group we obtain a series of probability measures for each term to belong to each variety in the group. Concretely, we calculate tf.idf weights for each term in the training set. With each weight, we calculate the probability as the relation between the sum of weights of the term belonging to the variety and the total sum of all its weights. In the end, we have the probability for each term to belong to each different variety of the language group. These probabilities were obtained from the training set. We must highlight that we learned a classifier for each language group, separately. Hence, the probabilities were computed locally for each group.

Once the language group of a new document is determined, to represent that document all its terms are computed with the previous probabilities for each language variety of such group. Then, we obtain six different measures from the computation of these probabilities: *1*) *Average*, computed as the sum of probabilities divided by the number of terms in the document; *2*) *Standard deviation*, computed as the root square of the sum of all probabilities minus the average; *3*) *Minimum probability*, the minimum of all probabilities computed for the document; *4*) *Maximum probability*, the maximum of all probabilities computed for the document; *5*) *Overall probability*, computed as the sum of all probabilities divided by the number of terms in the document and *6*) *Ratio*, computed as

the number of terms appearing in the document divided by the number of terms in the vocabulary. We obtain these 6 measures for each variety. Hence, we represent each document with a total of 6 features (described above), multiplied by the number of languages/varieties of its detected group. For each group, we used a Bayesian Net classifier as machine learning method.

The whole process is as follows. To predict the language of a new text, first we detect its language with the `ldig` language detector. Once we know the language group, we calculate the six aforementioned measures for each variety in the group and predict the variety with a Bayesian Net classifier.

For the close submission we represented all varieties together. This implies to learn all the probabilities together and then to predict the right variety with a single multi-class classifier. In this case, we represent each document with a total of 84 features: the 6 features described above, multiplied by 14 languages/varieties of the task. As classification method, we used Naive Bayes due to performance issues in training phase.

3 Experimental Results

In this section we show the evaluation of the proposed methodology when participating in the DSL 2015 shared task. Firstly, the dataset and the evaluation methodology are described. Then, the official results are shown. We detected a bug that is also described in this section. Finally, we explain our participation in the open and close submissions respectively, and discuss a comparison between both submissions.

3.1 Dataset and Methodology

We used the DSLCC v.2.0 (Tan et al., 2014) dataset. The dataset contains sentences extracted from news in different languages and dialects. Table 1 summarises the different languages and varieties contained in the dataset. The group coded as *xx* is built with sentences of different languages.

The length of each sentence ranges from 20 to 100 tokens. For each language or dialect, this dataset contains 18.000 instances for training, 2.000 instances for development and 1.000 instances for each test set. A summary of the total number of instances is shown in Table 2. The dataset is composed of two test sets, A and B. They both contain the same instances, but the test B was processed with a Named Entity Recogniser (NER) to replace Named Entities (NE) by placeholders. This set is named NE blinded.

Group	Language	Code
South-Eastern Slavic	Bulgarian	bg
	Macedonian	mk
Spanish	Argentinian	es-AR
	Peninsular	es-ES
Portuguese	Brazilian	pt-BR
	European	pt-PT
South-Western Slavic	Bosnian	bs
	Croatian	hr
	Serbian	sr
Austranesian	Indonesian	id
	Malay	my
West Slavic	Czech	cz
	Slovak	sk
Other		xx

Table 1: Languages in the DSLCC v.2.0 dataset.

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

Table 2: Number of instances per set.

We used the training set to learn probabilities and the corresponding machine learning models. We tested our methods with the development set using the Weka GUI[7] (Witten and Frank, 2005). We built a Java application to predict documents in the test set by using the models previously learned with Weka. In the following sections we explain the specific approach for both open and close submissions. We present comparative results among development, test A and test B. We also carried out a statistical significance test between results for both test sets. We used the following notation for confidence levels: * at 95% and ** at 99%

3.2 Task Results and Software Bug

Our results at the DSL task are shown in Table 3.

Open		Close	
Test A	Test B	Test A	Test B
91.84	89.56	64.04	62.78

Table 3: Identification accuracies for the open and close submission for tests A and B.

We detected a drop of accuracies between test and development. We reproduced the drop of accuracies by comparing results obtained with the

[7] http://www.cs.waikato.ac.nz/ml/weka/

Weka GUI and with our Java application in the development set. In Table 4 accuracies obtained by both methods are shown. The bug was present in our Java application. We did not compute properly the probabilities for the input set. Furthermore, some features were considered in wrong order.

Language	Weka GUI	Java App
es	87.75	86.60
pt	89.35	88.58
hr	83.63	79.96
id	99.43	99.42
all together	86.08	63.21

Table 4: Performance differences between Weka GUI and our Java application in the development set.

We could not fix the bug before submission time, and therefore our final results were much lower than we have expected. This is especially important in the close submission where the accuracy dropped more than 20%. In the following sections we analyse results with the error fixed.

3.3 Open Submission

We approached the open submission as a two-step process. Firstly, we used the `ldig` language detector to obtain the language group. The `ldig` detector was trained from the xml Wikipedia abstracts. We do not explicitly set any language group. Instead, the `ldig` language detector detects similar languages/dialects as a single language. We profit this fact to establish the language groups. The accuracy of this step for the development set is shown in Table 5.

Languages/Varieties	Language Group	Accuracy
bg	bg	99.80
mk	mk	100.00
es-AR, es-ES	es	99.96
pt-BR, pt-PT	pt	99.72
hr, bs, sr	hr	99.73
id, my	id	99.92
cz	cz	99.63
sk	sk	99.65
other languages	xx	99.90
	overall	99.81

Table 5: Identification accuracies of the `ldig` language detector in the development set.

In this step, we could detect Bulgarian (*bg*), Czech (*cz*), Macedonian (*mk*) and Slovak (*sk*). With respect to the other varieties, they were detected as follows: South-Western Slavic languages (Croatian, Bosnian and Serbian) were detected as Croatian (*hr*); Austranesian languages (Indonesian and Malay) were detected as Indonesian (*id*); and Spanish languages (Peninsular and Argentinian) and Portuguese languages (European and Brazilian) as their respective groups (*es* and *pt*). We classified as *xx* all the rest. Once the language group was identified, we applied our probabilistic method to detect the corresponding variety. Results for the development, test and NE blinded test sets are shown in Table 6.

Language	Accuracy		
	Devel.	Test A	Test B
bg*	99.80	99.90	99.80
mk*	100.00	99.90	100.00
es-ES	88.00	84.70	79.50
es-AR*	87.50	88.00	87.70
pt-PT	88.60	87.40	94.00
pt-BR	90.10	90.03	68.50
bs*	78.35	78.00	74.40
hr*	86.15	85.80	85.40
sr**	86.40	86.40	82.70
id	99.40	99.40	92.90
my*	99.45	99.20	99.50
cz*	99.70	99.80	99.40
sk*	99.60	99.30	99.60
xx*	99.90	99.90	99.70
overall	93.07	92.71	90.22

Table 6: Identification accuracies for the open submission for development, test, and NE blinded test.

Results for groups with only one language (*bg*, *mk*, *cz*, *sk*) show accuracies over 99% for both development and test sets. Accuracies for groups with more than one variety are quite lower. But this is not the case of Austronesian (*id*) where the achieved results are greater than 99% except for the *id* variety in the NE blinded test. The worst results were obtained for South-Western Slavic (*hr*) where the classifier should discriminate among three classes. The significance test shows us that our method is quite robust against blinded Named Entities in case of South-Western Slavic varieties (*bs*, *hr* and *sr*), Malay (*my*) and Argentinian Spanish (*es-AR*).

3.4 Close Submission

In the close submission we trained from the whole training set a multi-class classifier for the set of 14 different languages. The results are summarised in Table 7.

We can see that overall results for test B (72.11%) are much lower than for test A (85.57%) and development (86.08%). In this line, results for most languages are significantly different, except

Language	Accuracy		
	Devel.	Test A	Test B
bg	98.15	97.50	95.10
mk*	98.95	98.20	98.20
es-ES	87.55	84.80	48.70
es-AR**	67.05	70.00	74.10
pt-PT	82.15	81.20	58.30
pt-BR	72.45	72.50	65.90
bs	55.70	54.30	86.20
hr	80.85	78.88	13.10
sr	74.40	74.70	7.80
id	97.75	97.60	92.00
my	94.25	93.60	97.60
cz	98.45	98.40	94.40
sk	98.80	97.60	79.30
xx*	98.55	98.50	98.80
overall	86.08	85.57	72.11

Table 7: Identification accuracies for the close submission for development, test, and NE blinded test.

for the Argentinian Spanish (*es-AR*), Macedonian (*mk*) and Other (*xx*) groups. This may be due to the probabilities of terms corresponding to NE, which may cause confusion between some varieties.

3.5 Comparison between Methods

In Table 8, the comparative results between open and close approaches in the development set are shown. It is noteworthy that both approaches obtained lower results with the same groups (*es*, *pt* and *hr*). Regarding groups with only one language (*bg*, *mk*, *cz* and *sk*), both approaches obtained accuracies over 95%. We carried out the significance test but we cannot assert that any system performs equal for both open and close submissions. Therefore, we can conclude that the two-step method for the open submission was more accurate than dealing with all the varieties together.

Group	Accuracy	
	Open	Close
bg	99.80	98.15
mk	100.0	98.95
es	87.75	77.30
pt	89.35	77.30
hr	83.63	70.32
id	99.43	96.0
cz	99.70	98.45
sk	99.60	98.80
xx	99.90	98.55
overall	93.07	86.08

Table 8: Identification accuracies for the open and close submissions in development set.

4 Conclusions

In this work we presented the NLEL_UPV_Autoritas team participation at the DSL shared task. We submitted runs for both open and close tasks, for both normal and NE blinded tests. For the open submission, we developed a two-step system: in the first step we detected the language group and then the specific variety. For the close submission, we approached the task as a multi-class classification problem with all the varieties together.

We detected a software bug that dropped our results significantly in the testing phase. We fixed the bug and presented comparative results among development, test A and test B. We can conclude that approaching the task in two steps allows for obtaining better results than identifying all varieties together. Other teams approached the DSL 2014 shared task with two-step classification systems, obtaining good results. In this vein, Goutte et al. (2014) obtained the highest overall accuracy (95.71%) by predicting first the language group with a probabilistic generative classifier, and then predicting the variety within that group with a voting combination of classifiers. Porta and Sancho (2014) also predicted first the group and then the variety, with a hierarchical classifier based on maximum-entropy classifiers. They obtained an overall accuracy of 92.6%. Regarding varieties, the hardest prediction came with South-Western Slavic language, followed by Spanish and Portuguese. The Austranesian group was properly identified with both approaches. Groups composed by only one language obtained higher accuracies both in open and close approaches.

As future work we plan to approach the task implementing our own language detector. Moreover, we would like to investigate how to improve the accuracy in more similar languages than South-Western Slavic, Spanish or Portuguese, and to better deal with Named Entities.

Acknowledgement

The work of the second author was partially funded by Autoritas Consulting SA and by Spanish Ministry of Economics under grant ECOPORTUNITY IPT-2012-1220-430000. The work of the third author has been carried out within the framework of the European Commission WIQEI IRSES (no. 269180) and DIANA - Finding Hidden Knowledge in Texts (TIN2012-38603-C02)

projects, and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

References

Simon Carter, Wouter Weerkamp, and Manos Tsagkias. 2013. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47(1):195–215.

Marc Franco-Salvador, Francisco Rangel, Paolo Rosso, Mariona Taulé, and M. Antònia Martí. 2015. Language variety identification using distributed representations of words and documents. In *Proceeding of the 6th International Conference of CLEF on Experimental IR meets Multilinguality, Multimodality, and Interaction (CLEF 2015)*, volume LNCS(9283). Springer-Verlag.

Cyril Goutte, Serge Léger, and Marine Carpuat. 2014. The nrc system for discriminating similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 139–145, Dublin, Ireland, August. Association for Computational Linguistics.

Ben King, Dragomir Radev, and Steven Abney. 2014. Experiments in sentence language identification with groups of similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 146–154, Dublin, Ireland, August. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

Marco Lui, Ned Letcher, Oliver Adams, Long Duong, Paul Cook, and Timothy Baldwin. 2014. Exploring methods and resources for discriminating similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 129–138, Dublin, Ireland, August. Association for Computational Linguistics.

Wolfgang Maier and Carlos Gómez-Rodrıguez. 2014. Language variety identification in spanish tweets. *LT4CloseLang 2014*, page 25.

Jordi Porta and José-Luis Sancho. 2014. Using maximum entropy models to discriminate between similar languages and varieties. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 120–128, Dublin, Ireland, August. Association for Computational Linguistics.

Matthew Purver. 2014. A simple baseline for discriminating similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 155–160, Dublin, Ireland, August. Association for Computational Linguistics.

Francisco Rangel and Paolo Rosso. 2015a. On the impact of emotions on author profiling. *Information Processing & Management*, (In press) doi: 10.1016/j.ipm.2015.06.003.

Francisco Rangel and Paolo Rosso. 2015b. On the multilingual and genre robustness of emographs for author profiling in social media. In *Proceeding of the 6th International Conference of CLEF on Experimental IR meets Multilinguality, Multimodality, and Interaction (CLEF 2015)*, volume LNCS(9283). Springer-Verlag.

Francisco Rangel, Paolo Rosso, Moshe Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the author profiling task at pan 2013. In *Forner P., Navigli R., Tufis D.(Eds.), Notebook Papers of CLEF 2013 LABs and Workshops. CEUR-WS.org, vol. 1179*.

Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. 2014. Overview of the 2nd author profiling task at pan 2014. In *Cappellato L., Ferro N., Halvey M., Kraaij W. (Eds.) CLEF 2014 Labs and Workshops, Notebook Papers. CEUR-WS.org, vol. 1180*.

Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd author profiling task at pan 2015. In *Cappellato L., Ferro N., Gareth J. and San Juan E. (Eds). CLEF 2015 Labs and Workshops, Notebook Papers. CEUR-WS.org*.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic identification of arabic language varieties and dialects in social media. *SocialNLP 2014*, page 22.

Nakatani Shuyo. 2010. Language detection library for java. http://code.google.com/p/language-detection/.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora*, pages 11–15, Reykjavik, Iceland.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57.

Ian H Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Marcos Zampieri and Binyam Gebrekidan Gebre. 2012. Automatic identification of language varieties: The case of portuguese. In *KONVENS2012-The 11th Conference on Natural Language Processing*, pages 233–237. Österreichischen Gesellschaft für Artificial Intelligende (ÖGAI).

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the dsl shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland, August. Association for Computational Linguistics.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the dsl shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial)*, Hissar, Bulgaria.