The GW/LT3 VarDial 2016 Shared Task System for Dialects and Similar Languages Detection

Ayah Zirikly
George Washington University
Washington, DC
ayaz@gwu.edu

Bart Desmet
LT3, Ghent University
Ghent, Belgium
bart.desmet@ugent.be

Mona Diab
George Washington University
Washington, DC
mtdiab@gwu.edu

Abstract

This paper describes the GW/LT3 contribution to the 2016 VarDial shared task on the identification of similar languages (task 1) and Arabic dialects (task 2). For both tasks, we experimented with Logistic Regression and Neural Network classifiers in isolation. Additionally, we implemented a cascaded classifier that consists of coarse and fine-grained classifiers (task 1) and a classifier ensemble with majority voting for task 2. The submitted systems obtained state-of-the-art performance and ranked first for the evaluation on social media data (test sets B1 and B2 for task 1), with a maximum weighted F1 score of 91.94%.

1 Introduction

The 2016 DSL shared task objective was to correctly identify the different variations of similar languages (Malmasi et al., 2016). DSL2016 covered two main subtasks:

- Task 1: discriminating between similar languages from the same language family and between national language varieties. Covered languages and varieties are:
  - I Bosnian (bs), Croatian (hr) and Serbian (sr) from the South Slavic language family
  - II Malay (my) and Indonesian (id) from the Austronesian language family
  - III Portuguese from Brazil (pt-BR) and Portugal (pt-PT)
  - IV Spanish from Argentina (es-AR), Mexico (es-MX) and Spain (es-ES)
  - V French from France (fr-FR) and Canada (fr-CA)
- Task 2: Arabic dialect identification. The task includes Modern Standard Arabic (MSA) and the Egyptian (EGY), Gulf (GLF), Levantine (LAV) and North African (NOR) dialects.

Both tasks were evaluated in two tracks: closed (no external resources or additional training data are allowed) and open. The shared task involves predicting different languages for groups I and II from Task 1, identifying different variants of the same language in groups III, IV, V from Task 1, and predicting dialects in Task 2. Furthermore, Task 1 was evaluated on in-domain and out-of-domain test sets.

The experimental approaches described in this paper include preprocessing methods to prepare the data, feature engineering, various machine learning methods (Logistic Regression, Support Vector Machines and Neural Networks) and system architectures (one-stage, two-stage and ensemble classifiers). Additionally, we collected Twitter training data for Task 2 and studied its impact on prediction performance. GW/LT3 participated in Task 1 *(closed)* and Task 2 *(closed and open)*.

The rest of the paper is organized as follows: Section 2 presents a brief overview of work in similar languages identification and previous DSL tasks. Section 3 describes the overall methodology, whereas Section 4 and 5 discuss the datasets, preprocessing, experimental results and analysis in detail for each task. Section 6 concludes this paper.

Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects,
pages 33–41, Osaka, Japan, December 12 2016.

2 Related Research

Language identification is an active field of research, where in recent years increased attention has been given to the identification of closely related languages, language variants and dialects, which are harder to distinguish. The three editions of the DSL shared task on detecting similar languages have provided a forum for benchmarking various approaches. For a detailed overview of the previous editions and their related work, we refer to the overview papers of Zampieri et al. (2014) and Zampieri et al. (2015).

State-of-the-art approaches to related language identification rely heavily on word and character n-gram representations. Other features include the use of blacklists and whitelists, language models, POS tag distributions and language-specific orthographical conventions (Bali, 2006; Zampieri and Gebre, 2012). For systems, a wide range of machine learning algorithms have been applied (Naive Bayes and SVM classifiers in particular), with work on optimization and dimensionality reduction (Goutte et al., 2014), and on ensembling and cascading, which yielded the best-performing systems in the 2015 edition (Goutte and Léger, 2015; Malmasi and Dras, 2015).

Previous approaches for Arabic dialect detection, a new task introduced in this shared task edition, use similar approaches. Sadat et al. (2014) argue that character n-gram models are well suited for dialect identification tasks because most of the variation is based on affixation, which can be easily modeled at the character level.

Also new to this edition of the shared task is the evaluation on social media data. In 2014, the Tweet-LID shared task specifically addressed the problem of language identification in very short texts (Zubiaga et al., 2014). This brought to light some of the challenges inherent to the genre: a need for a better external resources to train systems, low accuracy on underrepresented languages and the inability to identify multilingual tweets.

3 System Description

We experiment with a number of machine learning methods that range from conventional methods such as Logistic Regression to Deep Learning.

Feature Set We experimented with a simple feature set similar to those that proved effective in previous DSL tasks (Goutte and Léger, 2015). We employ word and character n-gram representations as features in the closed submission for Task 1. Additionally, we incorporate lexical features based on Arabic dialect dictionaries. We generated `GLF`, `EGY`, `LAV`, and `NOR` noisy dictionaries that are collected from Twitter where a filter based on the geolocation field from Twitter API is applied to reflect the targeted dialects (e.g. `KW` $\rightarrow$ `GLF`). The `MSA` dictionary is based on the unique vocabulary set in Arabic Gigaword. The dictionary features are a set of 5 features (one per dialect) where each feature value represents the in-dictionary occurrence frequencies (e.g. *kdh mA ySH$* [EN: *This is not right*]: `GLF_dic:1, EGY_dic:3, MSA_dic:1, LAV_dic:1, NOR_dic:1`).

Classifiers

Support Vector Machines (SVM): we experimented with SVMs and found that it produces worse results in comparison to other classifiers. As a result, we did not submit a run that implements SVM.

Logistic Regression (LR) classifier: the intuition behind using LR as opposed to Support Vector Machines (SVM) is that LR works better in scenarios where the classes are close to each other and when the predictors can near-certainly determine the output label. We use LR for both Task 1 and Task 2 as one of the submitted runs, where LR produces state-of-the-art results for Task 1 on out-of-domain data. All LRs are trained with L2 regularization and a cost C of 1.

Neural Network (NN) classifier: we also experiment with NNs, because they have proven effective in modelling a wide range of complex NLP tasks. All NNs are trained with a single hidden layer of 500 neurons, using softmax activation and Adaptive Moment Estimation (Adam) to optimize the stochastic gradient descent.

Two-stage classifier: for Task 1, we implemented a two-stage classifier where we first train a system to predict the coarse-grained language group class. Then, for every language group we built a model

that predicts the fine-grained variant class. A detailed description of this classifier is depicted in Figure 1. **Ensemble with majority voting:** for Task 2, we implemented an ensemble classifier that takes

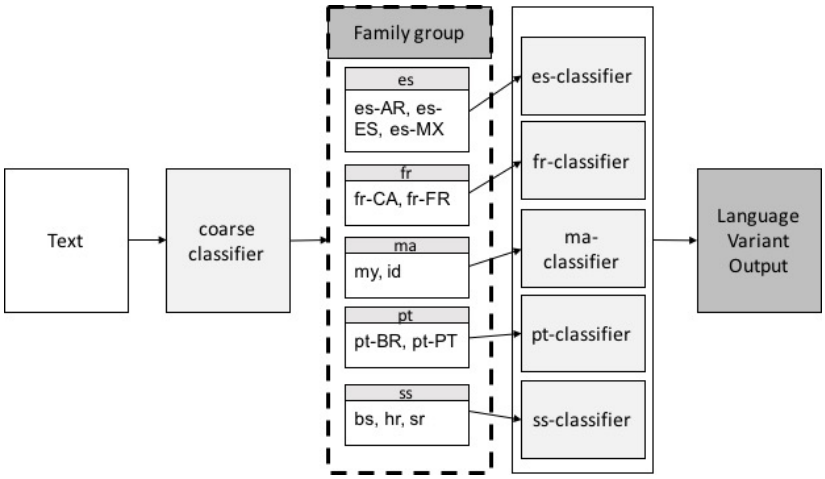


Figure 1: Two-stage coarse-fine classifier

the majority vote of 1 LR and 2 NN classifiers' output and produces the majority label. Ties are broken by taking the output of the best-performing individual classifier. The number and selection of classifiers participating in the ensemble was determined experimentally on the development data.

4 Task 1

Task 1 focuses on predicting the correct label of language variant among classes of similar languages and variants.

4.1 Datasets

The dataset provided contains:

- Training and development data: a balanced train and dev set containing fragments from newswire text (18000 training and 2000 dev instances per class)
- Test data: class-balanced in-domain (test set A: 12000 instances) and out-of-domain data collected from social media (test sets B1 and B2 with 500 instances each, pertaining to the South Slavic and `pt` families)

4.2 Preprocessing

In order to reduce data dimensionality and improve lexical recall, preprocessing was applied to the datasets. This was especially relevant for the out-of-domain datasets B1 and B2, which were noisy in nature since they had been collected from Twitter. We performed the following normalization operations:

- number masking (e.g. $1990 \Rightarrow 8888$)
- URL replacement (e.g. *ttg.uni-saarland.de/vardial2016* $\Rightarrow URL$)
- in words elongated with character flooding, repetitions are limited to two occurrences (e.g. $goooood \rightarrow good$)
- removal of at-mentions, retweets and HTML tags
- lowercasing of all text

Additionally, we applied language filtering on the tweets in datasets B1 and B2. The task was to determine the primary language variant of a Twitter user, given a collection of his or her tweets. However, Twitter users do not consistently tweet in the same language: some tweets may be in a different language entirely, and some may have internal code switching. Because such tweets can confuse a classifier, we removed all tweets that could not be confidently assigned to one of the language groups under study. We used the probability outputs of a NN coarse-grained classifier to remove all tweets that had less than 95% of the probability mass concentrated in one category.

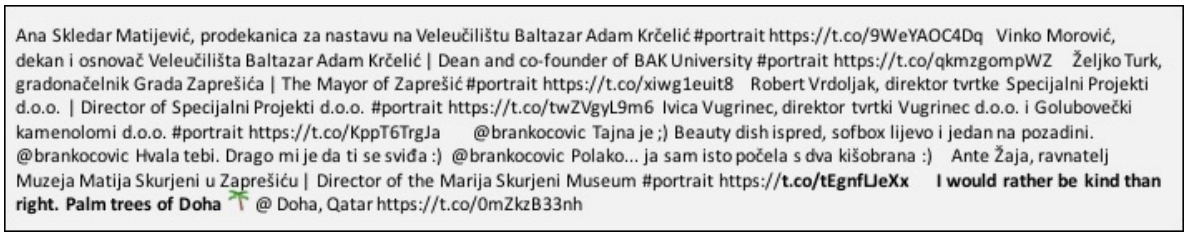

Ana Skledar Matijević, prodekanica za nastavu na Veleučilištu Baltazar Adam Krčelić #portrait https://t.co/9WeYAOC4Dq Vinko Morović, dekan i osnovač Veleučilišta Baltazar Adam Krčelić | Dean and co-founder of BAK University #portrait https://t.co/qkmzgompWZ Željko Turk, gradonačelnik Grada Zaprešića | The Mayor of Zaprešić #portrait https://t.co/xiwg1euit8 Robert Vrdoljak, direktor tvrtke Specijalni Projekti d.o.o. | Director of Specijalni Projekti d.o.o. #portrait https://t.co/twZVgyL9m6 Ivica Vugrinec, direktor tvrtki Vugrinec d.o.o. i Golubovečki kamenolomi d.o.o. #portrait https://t.co/KppT6TrgJa @brankocovic Tajna je ;) Beauty dish ispred, sofbox lijevo i jedan na pozadini. @brankocovic Hvala tebi. Drago mi je da ti se sviđa :) @brankocovic Polako... ja sam isto počela s dva kišobrana :) Ante Žaja, ravnatelj Muzeja Matija Skurjeni u Zaprešiću | Director of the Marija Skurjeni Museum #portrait https://**t.co/tEgnfLJeXx** **I would rather be kind than right. Palm trees of Doha** @ Doha, Qatar https://t.co/0mZkzB33nh

Figure 2: Example of out-of-domain dataset entry

4.3 Postprocessing

For the B1 and B2 test sets, which only contain 2 of the 5 language groups, we normalize predictions pertaining to an incorrect language group by backing off to the highest-probability available class. In the case of the cascaded classifier, this is done in the first stage.

4.4 Results

The GW/LT3 team submitted to the closed track for Task 1, where no external training data or resources could be used. For each dataset, three systems were submitted (as explained in Section 3), with the following settings:

- LR: character (2-6) and word n-grams (1-3) with term-frequency weighting
- NN: binary character n-gram features (2-6), 35 epochs of training
- Cascade: both the coarse (language group) and fine-grained classifier use LR, with the same feature set as described above for LR

GW/LT3 ranked **first** in the out-of-domain evaluation (test sets B1&B2) and **third** for in-domain test set A. As shown in Table 1, the LR classifier yields the best performance on the B1 and B2 test sets, with an accuracy of 92.0% and 87.8%, respectively. It is narrowly beaten by the cascaded approach on test set A (88.7%).

The state-of-the-art performance on the B1 and B2 test sets may indicate that adequate preprocessing is a prerequisite when dealing with noisy social media data. Both the normalization steps and the aggressive filtering of code-switched tweets based on language family detection may have been effective for improving performance over competing systems.

Data / Metric	A			B1			B2		
	LR	NN	2-stage[3]	**LR[1]**	NN	2-stage	**LR[1]**	NN	2-stage
Accuracy	88.59	85.02	**88.70**	**92.00**	89.60	91.20	**87.80**	86.00	87.20
F1-weighted	88.60	84.93	**88.70**	**91.94**	89.45	91.12	**87.73**	85.81	87.13

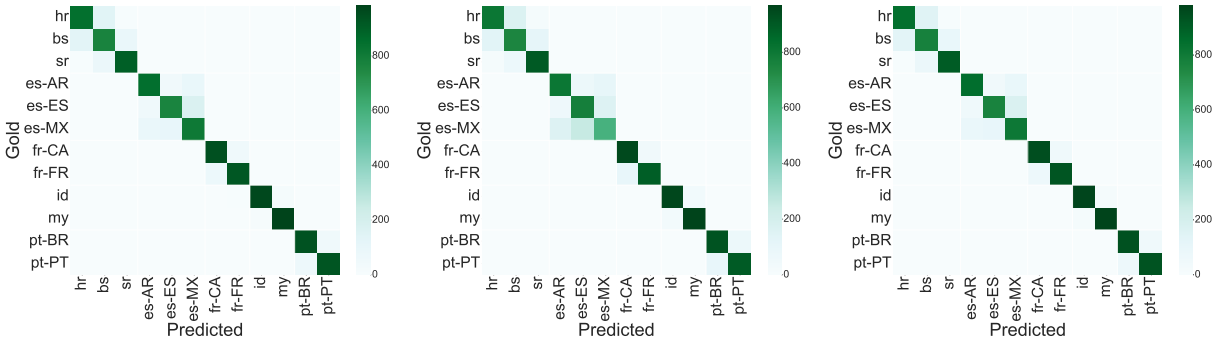
Table 1: Task 1 results. System ranks are indicated in superscript.

Based on the confusion matrices for the in-domain dataset, we note a very similar behavior among the three different approaches, especially LR & two-stage. We note that NN consistently performs worse than the other two approaches with a marked accuracy degradation in the more closely language variants,

Method \ Variant	hr	bs	sr	es-ar	es-es	es-mx	fr-ca	fr-fr	id	my	pt-br	pt-pt
	A											
LR	85	77	90	85	80	77	94	93	98	98	93	93
NN	82	75	88	79	73	63	92	91	96	96	91	91
2-stage	85	77	90	85	80	78	94	93	98	98	94	93
	B1											
LR	93	86	92	-	-	-	-	-	-	-	94	94
NN	88	82	95	-	-	-	-	-		-	92	91
2-stage	93	86	92	-	-	-	-	-	-	-	93	92
	B2											
LR	92	85	91	-	-	-	-	-	-	-	86	84
NN	90	80	92	-	-	-	-	-	-	-	85	82
2-stage	92	84	91	-	-	-	-	-	-	-	85	83

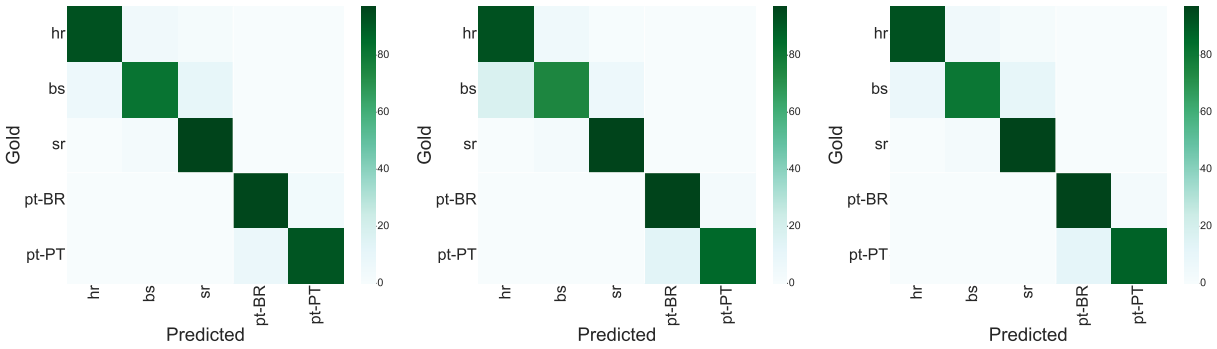
Table 2: Task 1 per-variant F1-score

such as the Portuguese and Spanish language groups. The NN approach performs notably poorly for the detection of Mexican Spanish with a recall of 58% in comparison to 81% for LR. However, it is worth noting that performance for Mexican Spanish is poor across classifiers (Table 2). Together with Bosnian (across datasets), it appears to be harder to predict than other language variants.



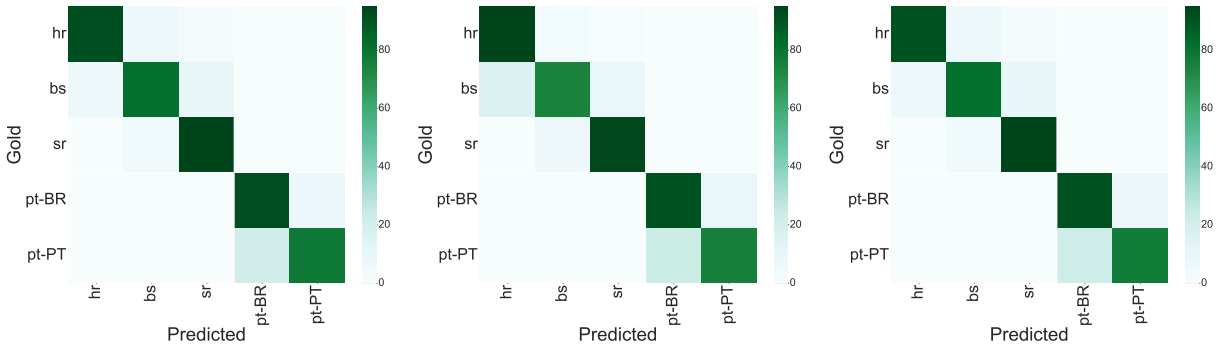
(a) LR (b) NN (c) 2-stage

Figure 3: A Confusion Matrices



(a) LR (b) NN (c) 2-stage

Figure 4: B1 Confusion Matrices



(a) LR (b) NN (c) 2-stage

Figure 5: B2 Confusion Matrices

5 Task 2

Task 2 aims to predict the correct Arabic dialect from a set of 5 different dialects. GW/LT3 submitted systems to both the open and closed for the Arabic subtracks.

5.1 Datasets

The dataset (Ali et al., 2016) provided contains Automatic Speech Recognition (ASR) transcripts in Buckwalter encoding[1] and is divided into:

- Training data: unlike Task 1, the training data is unbalanced and contains 1578 `EGY`, 1672 `GLF`, 1758 `LAV`, 999 `MSA`, and 1612 `NOR` instances (total of 7619)
- Test data: ASR transcripts containing 315 `EGY`, 256 `GLF`, 344 `LAV`, 274 `MSA`, and 351 `NOR` instances (total of 1540)

External datasets For the open submission, we used dialect dictionaries to make in-vocabulary frequency count features (as explained in 3). For MSA, we used the Arabic Gigaword vocabulary, whereas for other dialects we built dictionaries based on data collected from Twitter. We are aware that using social media data invariably introduces noise, both in terms of misspelled vocabulary entries and with relation to incorrect geographical information. However, as argued by Mubarak and Darwish (2014), such information still provides acceptable dialectal corpora. We filtered the collected tweets based on the countries of interest that map to the targeted dialects of the shared task (e.g. $Syria \rightarrow LAV$). Before creating the dictionaries, we apply normalization (hamza normalization, emoji and URL removal, ...). The resulting dictionary sizes were 76,721 for `GLF`, 22,003 for `EGY`, 10,000 for `LAV`, 286,559 for `MSA` and 6,343 for `NOR`.

5.2 Preprocessing

We tested applying letter normalization during the train/dev phase, where we normalized the different shapes of hamza $(', |, >, \&, <, ,)$ to Alif (A). However, we noted that this type of normalization did not improve performance, which is why it was omitted in the final systems. However, preprocessing on the dictionaries collected from Twitter was applied in a similar fashion as the one described in 4.2.

5.3 Results

Settings of the 3 submitted runs for both tracks (as explained in Section 3), were as follows:

- LR: character (2-6) and word n-grams (1-3) without term-frequency weighting, additional dictionary features for the open track
- NN: binary character n-gram features (2-6), 35 epochs of training

[1] http://www.qamus.org/transliteration.htm

- Ensemble: 1 NN classifier with character (3-5) and word (1) n-grams, 1 NN classifier with character n-grams (2-6) and 1 LR classifier with character n-grams (1-6) with MSA dictionary features for the open track

GW/LT3 ranks 2nd and 5th in the open and closed settings respectively, using the ensemble approach (EMV) described in Section 3. Table 3 shows the three submitted runs' performance under the closed and open settings. We note that adding extra features using the external resources, or even adding them as extra training data during the train/dev phase, did not improve the performance of the systems. This can likely be explained by limited overlap in genre between the training and test data and the Twitter data. In Table 4, we note that EMV produces the best performance per dialect, with `MSA` being the easiest dialect to identify. This may be explained by the fact that `MSA` is highly distinguishable from other dialects, as opposed to the high overlap between dialects' vocabularies.

Data / Metric	Closed			Open		
	LR	NN	EMV[5]	LR	NN	EMV[2]
Accuracy	44.42	49.03	**49.03**	44.35	49.03	**49.09**
F1-weighted	44.79	49.17	**49.22**	44.74	49.17	**49.29**

Table 3: Task 2 results. System ranks are indicated in superscript.

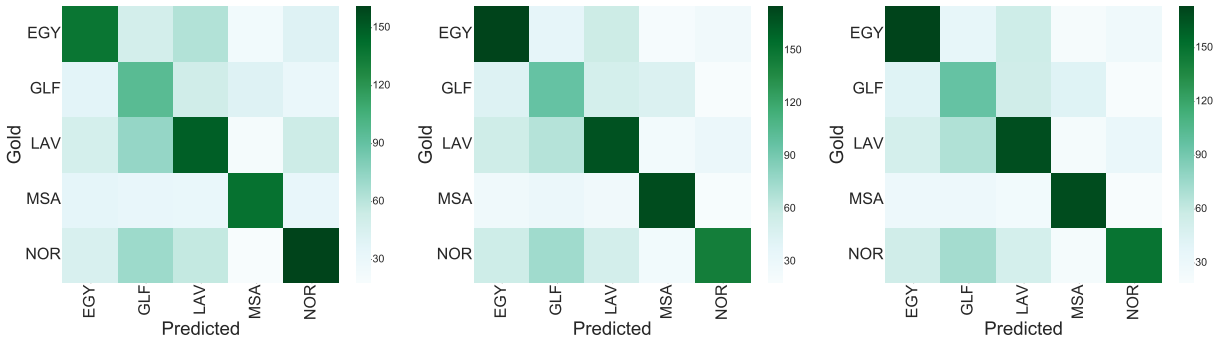
Dialect / Method	EGY	GLF	LAV	MSA	NOR
	CLOSED				
LR	45	33	43	54	48
NN	52	**35**	48	61	49
EMV	**52**	34	**48**	**61**	**50**
	OPEN				
LR	44	33	43	55	48
NN	52	35	48	61	49
EMV	**52**	**35**	**48**	**61**	**50**

Table 4: Task 2 dialects F1-score

Based on Figure 6 and 7, we note that our systems perform in a very similar behavior under the open and closed settings, which is due to the small number of added features under the open settings as opposed to the closed. `GLF` dialect represents the highest challenge for our systems with F1-score of 35% (as shown in Table 4). Based on the confusion matrix, we note that `GLF` is often mispredicted as `LAV` or `MSA`. Additionally, we note that `MSA` yields the best performance among the various dialects, a result aligning with the findings of Zaidan and Callison-Burch (2014). EMV produces the best overall accuracy and F-score results with a performance that is very close to the NN system, as two of the three votes belong to NN systems with different parameters.

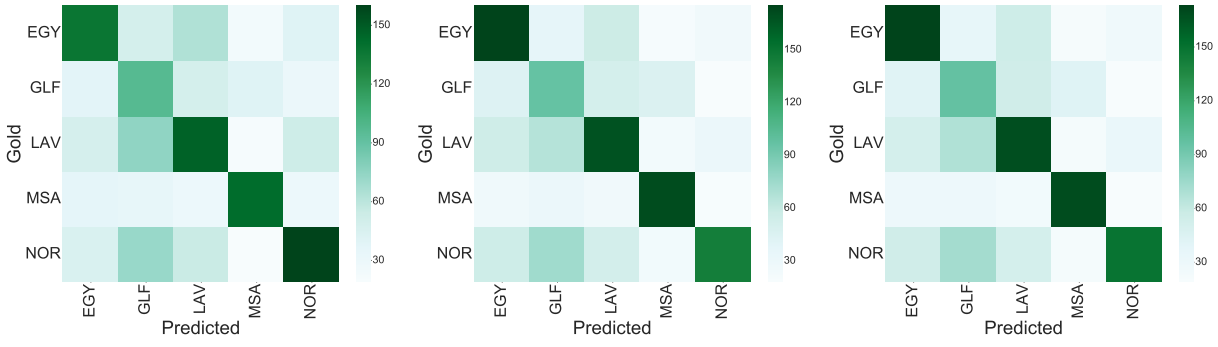
6 Conclusion & Future Work

In this paper, we discussed the collaborative work between George Washington University (GW) and Ghent University (LT3), where GW/LT3 submitted systems to both 2016 DSL Task 1 (closely related languages and variants) and Task 2 (Arabic dialect identification). The performance of our best run on out-of-domain data for Task 1 ranked first, using a Logistic Regression classifier. We hypothesize that adequate preprocessing of noisy social media data may be a prerequisite for good performance. Complex system architectures such as cascaded classification or ensembling did not yield significant improvements over the one-stage classifiers. Given the promising results of the single-layer Neural Networks for the Arabic dialect detection task, we intend to investigate alternative Deep Learning methodologies in future work.



(a) LR (b) NN (c) 2-stage

Figure 6: C Closed Confusion Matrices



(a) LR (b) NN (c) 2-stage

Figure 7: C Open Confusion Matrices

Acknowledgements

We would like to thank the organizers for an interesting shared task. The first and third author were partially funded by DARPA DEFT subcontract from Columbia University. The second author was funded by the Flemish government agency for Innovation by Science and Technology, through the AMiCA project (IWT SBO 120007).

References

Ahmed Ali, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell, and Steve Renals. 2016. Automatic dialect detection in Arabic broadcast speech. In *Interspeech 2016*, pages 2934–2938.

Ranaivo-Malançon Bali. 2006. Automatic identification of close languages–case study: Malay and Indonesian. *ECTI Transaction on Computer and Information Technology*, 2(2):126–133.

Cyril Goutte and Serge Léger. 2015. Experiments in discriminating similar languages. In *Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, page 78.

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.

Shervin Malmasi and Mark Dras. 2015. Language identification using classifier ensembles. In *Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, page 35.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešic, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Proceedings of the 3rd Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (VarDial), Osaka, Japan*.

Hamdy Mubarak and Kareem Darwish. 2014. Using Twitter to collect a multi-dialectal corpus of Arabic. *ANLP 2014*, page 1.

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

Omar F Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

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 Ljubešic, 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.

Marcos Zampieri, Liling Tan, Nikola Ljubešic, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL shared task 2015. In *Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, page 1.

Arkaitz Zubiaga, Inaki San Vicente, Pablo Gamallo, José Ramom Pichel Campos, Iñaki Alegría Loinaz, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno-Fernández. 2014. Overview of TweetLID: Tweet language identification at SEPLN 2014. In *TweetLID@ SEPLN*, pages 1–11.