Arabic Dialect Identification in Speech Transcripts

Shervin Malmasi[1,2] **Marcos Zampieri**[3]

[1] Harvard Medical School, Boston, MA 02115, USA
[2] Macquarie University, Sydney, NSW, Australia
[3] University of Cologne, Germany

shervin.malmasi@mq.edu.au, marcos.zampieri@uni-koeln.de

Abstract

In this paper we describe a system developed to identify a set of four regional Arabic dialects (Egyptian, Gulf, Levantine, North African) and Modern Standard Arabic (MSA) in a transcribed speech corpus. We competed under the team name MAZA in the Arabic Dialect Identification sub-task of the 2016 Discriminating between Similar Languages (DSL) shared task. Our system achieved an F1-score of 0.51 in the closed training track, ranking first among the 18 teams that participated in the sub-task. Our system utilizes a classifier ensemble with a set of linear models as base classifiers. We experimented with three different ensemble fusion strategies, with the mean probability approach providing the best performance.

1 Introduction

The interest in processing Arabic texts and speech data has grown substantially in the last decade.[1] Due to its intrinsic variation, research has been carried out not only on Modern Standard Arabic (MSA), but also on the various Arabic dialects spoken in North Africa and in the Middle East. Research in NLP and Arabic dialects includes, most notably, machine translation of Arabic dialects (Zbib et al., 2012), corpus compilation for Arabic dialects (Al-Sabbagh and Girju, 2012; Cotterell and Callison-Burch, 2014), parsing (Chiang et al., 2006), and Arabic dialect identification (Zaidan and Callison-Burch, 2014). The latter has become a vibrant research topic with several papers published in the last few years (Sadat et al., 2014; Malmasi et al., 2015).

In this paper we revisit the task of Arabic dialect identification proposing an ensemble method applied to a corpus of broadcast speeches transcribed from MSA and four Arabic dialects: Egyptian, Gulf, Levantine, and North African (Ali et al., 2016). The system competed in the Arabic dialect identification sub-task of the 2016 edition of the DSL shared task (Malmasi et al., 2016b)[2] under the team name MAZA. The system achieved very good performance and was ranked first among the 18 teams that participated in the closed submission track.

2 Related Work

There have been several studies published on Arabic dialect identification. Shoufan and Al-Ameri (2015) presents a survey on NLP methods for processing Arabic dialectal data with a comprehensive section on Arabic dialect identification.

Two studies on Arabic dialect identification use the Arabic online commentary dataset (Zaidan and Callison-Burch, 2011), namely the one by Elfardy and Diab (2013) and the one by Tillmann et al. (2014) who developed systems to discriminate between MSA and Egyptian Arabic at the sentence level. The first study reports results of 85.5% accuracy and the latter reports 89.1% accuracy using a linear SVM classifier.

This work is licenced under a Creative Commons Attribution 4.0 International License. License details:
http://creativecommons.org/licenses/by/4.0/

[1] See Habash (2010) for an overview on Arabic NLP.

[2] http://ttg.uni-saarland.de/vardial2016/dsl2016.html

Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects,
pages 106–113, Osaka, Japan, December 12 2016.

Malmasi et al. (2015) evaluates the performance of different methods and features to discriminate between MSA and five Arabic dialects: Egyptian, Jordanian, Palestinian, Syrian, and Tunisian using the Multidialectal Parallel Corpus of Arabic (MPCA) (Bouamor et al., 2014). Malmasi et al. (2015) report results of 74.0% accuracy using a meta-classifier. Darwish et al. (2014) identified important lexical, morphological, and syntactic features to discriminate between MSA and Egyptian Arabic *tweets* reporting 94.4% accuracy.

Using the same dataset as the DSL 2016 Arabic dialect identification sub-task, Ali et al. (2016) propose an SVM method to discriminate between MSA and dialectal Arabic achieving perfect performance. Ali et al. (2016) proposes the same method to identify the four aforementioned Arabic dialects and MSA and reports 59.2% accuracy.

The work on Arabic dialect identification is related to several studies published on computational methods to discriminate between pairs or groups of similar languages, language varieties and dialects. This includes South Slavic languages (Ljubešić et al., 2007), Portuguese varieties (Zampieri and Gebre, 2012), English varieties (Lui and Cook, 2013), Persian and Dari (Malmasi and Dras, 2015a), Romanian dialects (Ciobanu and Dinu, 2016), and the two editions of the DSL shared task organized in 2014 and 2015 which included several groups of closely-related languages and language varieties such as Bosnian, Croatian and Serbian, Bulgarian and Macedonian, Czech and Slovak, and Mexican and Peninsular Spanish (Zampieri et al., 2014; Zampieri et al., 2015).

3 Methods

3.1 Data

For the first time, the DSL challenge includes a sub-task on Arabic dialect identification. The data for this sub-task was provided by the DSL shared task organizers and it is described in the aforementioned study by Ali et al. (2016). The corpus contains transcribed speech from Egyptian (EGY), Gulf (GLF), Levantine (LAV), North African (NOR), and MSA.

The training corpus contains a total of 7,619 sentences. An additional unlabelled test set containing 1,540 sentences was released one month later for the official evaluation. A breakdown of the number of training sentences for each of these classes is listed in Table 1.

Dialect	Class	Sentences
Egyptian	EGY	1,578
Gulf	GLF	1,672
Levantine	LAV	1,758
Modern Standard	MSA	999
North African	NOR	1,612
Total		7,619

Table 1: The breakdown of the dialectal training data provided (Ali et al., 2016).

3.2 Approach

There have been various methods proposed for dialect identification in recent years. Given its success in previous work, we decided to use an ensemble classifier for our entry. We follow the methodology described by Malmasi and Dras (2015b): we extract a number of different feature types and train a single linear model using each feature type. We extract the following feature types, each of them used to train a single classification model:

- Character n-grams (n = 1–6): these substrings, depending on the order, can implicitly capture various sub-lexical features including single letters, phonemes, syllables, morphemes and suffixes. They could capture interesting inter-dialectal differences that generalize better than word n-grams.
- Word unigrams: entire words can capture lexical differences between dialects.

We did not perform any pre-processing[3] on the data prior to feature extraction. This was not needed as the data are machine-generated ASR transcripts.[4]

For our base classifier we utilize a linear Support Vector Machine (SVM). SVMs have proven to deliver very good performance in discriminating between language varieties and in other text classification problems, SVMs achieved first place in both the 2015 (Malmasi and Dras, 2015b) and 2014 (Goutte et al., 2014) editions of the DSL shared task.[5]

The best performing system in the 2015 edition of the DSL challenge (Malmasi and Dras, 2015b) used SVM ensembles evidencing the adequacy of this approach for the task of discriminating between similar languages and language varieties. In light of this, we decided to test three ensemble methods described next.

- **System 1 - Plurality Ensemble**
  In this system each classifier votes for a single class label. The votes are tallied and the label with the highest number[6] of votes wins. Ties are broken arbitrarily. This voting method is very simple and does not have any parameters to tune. An extensive analysis of this method and its theoretical underpinnings can be found in the work of (Kuncheva, 2004, p. 112). We submitted this system as run 1.

- **System 2 - Median Probability Ensemble**
  In this ensemble method the probabilities assigned to each class by each classifier are ordered, and the median probability for each label is selected. Among these, the label with the highest median is selected (Kittler et al., 1998). As with the mean probability combiner, which we describe in the next section, this method measures the central tendency of support for each label as a means of reaching a consensus decision. We submitted this system as run 2.

- **System 3 - Mean Probability Ensemble**
  The probability estimates for each class are added together and the class label with the highest average probability is the winner. An important aspect of using probability outputs in this way is that a classifier's support for the true class label is taken in to account, even when it is not the predicted label (*e.g.* it could have the second highest probability). This method has been shown to work well on a wide range of problems and, in general, it is considered to be simple, intuitive, stable (Kuncheva, 2014, p. 155) and resilient to estimation errors (Kittler et al., 1998) making it one of the most robust combiners discussed in the literature. We submitted this system as run 3.

4 Cross-validation Results

In this section we investigate the impact of three variables in the classification performance: the features used, the data, and the type of ensemble used in our system.

We used the training data provided by the shared task organizers and performance cross-validation experiments testing 1) the performance of each individual feature in dialect identification (described in Section 4.1); 2) the impact of the amount of training data on the classification performance (presented in Section 4.2); and 3) the accuracy of each proposed ensemble method (discussed in Section 4.3).

4.1 Feature Performance

We first report our cross-validation results on the training data. We began by testing individual feature types, with results displayed in Figure 1.

As expected we observe that most character n-grams outperform word features. Character 4-grams, 5-grams, and 6-grams obtained higher results than those obtained using word uni-grams. The best results were obtained with character 4-grams achieving 65.95% accuracy and character 5-grams obtaining 65.70% accuracy.

[3]For example, case folding or tokenization.

[4]The data was transliterated using the Buckwalter scheme: `http://www.qamus.org/transliteration.htm`

[5]See Goutte et al. (2016) for a comprehensive evaluation.

[6]This differs with a *majority* voting combiner where a label must obtain over 50% of the votes to win. However, the names are sometimes used interchangeably.

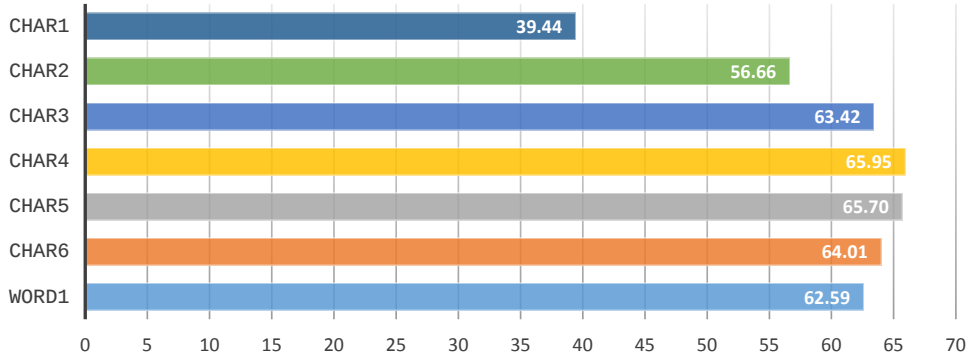

Figure 1: Cross-validation performance for each of our individual feature types.

4.2 Influence of Training Data

Next we look at the influence of the amount of training data in the Arabic dialect identification task. As the size of the training corpus provided by the shared task organizers is relatively small, we are interested in evaluating how this affects performance. A learning curve for a classifier trained on character 4-grams is shown in Figure 2. We observe that accuracy continues to increase, demonstrating potential for even better performance given a larger training corpus.[7]

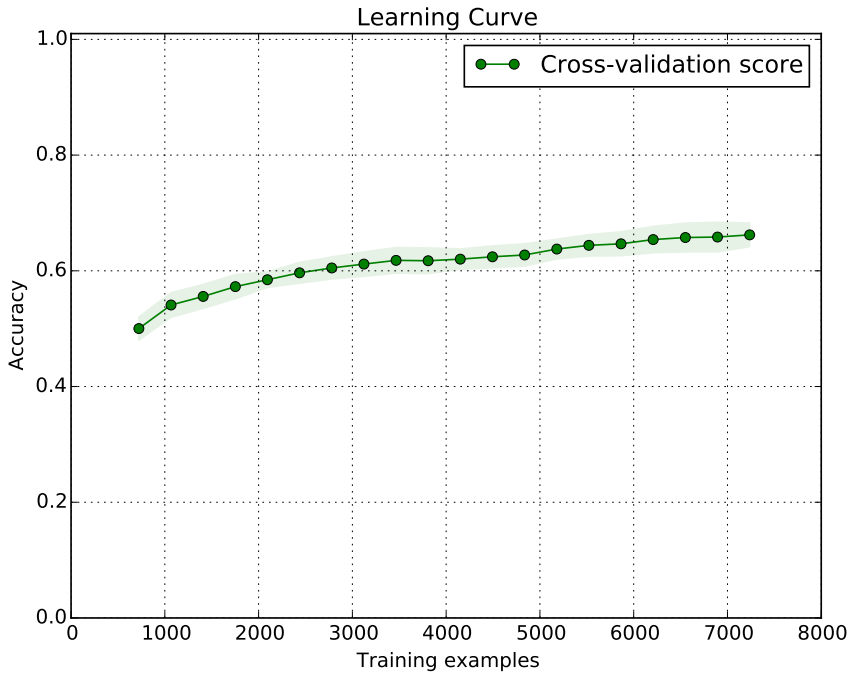

Figure 2: Learning curve for a classifier trained on character 4-grams using the training data.

4.3 Ensemble Methods

In this section we test our three ensemble configurations on the training data. Results are shown in Table 2. We note that all of the ensembles outperform individual features, with the mean probability combiner achieving the best result of 68%. For the voting ensemble, 344 of the 7619 samples (4.52%) resulted in ties which were broken arbitrarily.

[7]Due to lack of available comparable data, we only participated in the closed submission track.

System	Accuracy
Majority Class Baseline	0.2307
Voting Ensemble (System 1)	0.6755
Median Ensemble (System 2)	0.6782
Mean Probability Ensemble (System 3)	**0.6800**

Table 2: Cross-validation results for the Arabic training data.

5 Test Set Results

Finally, in this section we report the results of our three submissions generated from the unlabelled test data. The samples in the test set were slightly unbalanced with a majority class baseline of 22.79%. Shared task performance was evaluated and teams ranked according to the weighted F1-score which provides a balance between precision and recall. Accuracy, along with macro- and micro-averaged F1-scores were also reported.

Run	Accuracy	F1 (micro)	F1 (macro)	F1 (weighted)
Baseline	0.2279	—	—	—
System 1 (run1)	0.4916	0.4916	0.4888	0.4924
System 2 (run2)	0.4929	0.4929	0.4908	0.4937
System 3 (run3)	0.5117	0.5117	0.5088	**0.5132**

Table 3: Results for test set C (closed training).

Results for our three submissions are listed in Table 3. While Systems 1 and 2 achieved similar performance, System 3 outperformed them by approximately 2%, ranking first among the 18 teams who competed in the sub-task.

A confusion matrix for our best performing system is shown in Figure 3. We note that MSA is the most distinguishable dialect, while the Gulf dialect has the most misclassifications. Table 4 also shows per-class performance for our best system.

Class	Precision	Recall	F1-score	Sentences
EGY	0.50	0.56	0.53	315
GLF	0.33	0.36	0.35	256
LAV	0.51	0.48	0.49	344
MSA	0.60	0.63	0.61	274
NOR	0.62	0.52	0.56	351
Average/Total	0.52	0.51	0.51	1,540

Table 4: Per-class performance for our best system.

The results for all of our systems are much lower than the cross-validation results. This was a trend noted by other teams in the task. It is likely related to the sampling of the test set; it may have not been drawn from the same source as the training data.

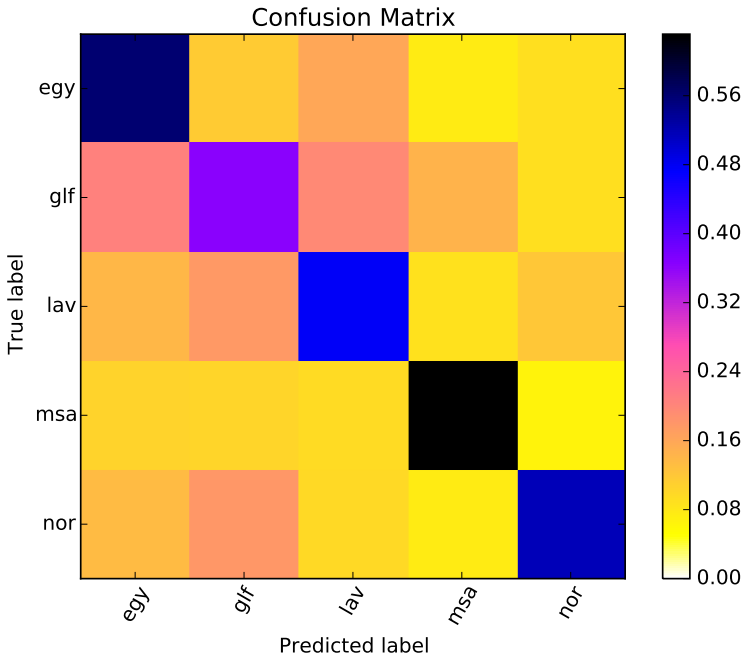

Figure 3: Confusion matrix for our top performing system on the test set.

5.1 Discussion

An important highlight from this work goes beyond Arabic dialect identification. Our work confirms the utility of ensemble methods for different text classification tasks. These methods have proven to perform well in similar shared tasks such as the recent Complex Word Identification (CWI) task at SemEval-2016 (Paetzold and Specia, 2016). A description of the ensemble system applied to CWI is presented in Malmasi et al. (2016a).

Regarding the task itself, this initial experiment shows that accurate dialect identification using ASR transcripts is not a trivial task. An interesting extension is the creation of joint audio-transcript classification models where transcript-based features like the ones used here are combined with acoustic features to capture phonological variation.

6 Conclusion

We presented three robust ensemble methods trained to discriminate between four Arabic dialects and MSA in speech transcripts. The best results were obtained by the Mean Probability Ensemble system (run 3) achieving 0.51 F1-score in the test data. The system outperformed all the 18 teams that participated in the Arabic dialect identification task of the DSL shared task 2016. A comprehensive overview of the 2016 DSL challenge including the results obtained by all participants is presented in Malmasi et al. (2016b).

Our paper also discusses two important variables in Arabic dialect identification, namely the performance of individual character- and word-based features for this task, highlighting that character 4-grams were the features which performed best using this dataset, and the influence of the amount of training data in the classifiers' performance.

As discussed in Section 2, Arabic dialect identification methods are related to methods developed to discriminate between similar languages and language varieties. In future work we would like to evaluate whether our system also achieves good performance discriminating between the languages and language varieties available in the DSL corpus collection (DSLCC) (Tan et al., 2014).

References

Rania Al-Sabbagh and Roxana Girju. 2012. YADAC: Yet another Dialectal Arabic Corpus. In *Proceedings of LREC*.

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 *Proceedings of Interspeech*.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A Multidialectal Parallel Corpus of Arabic. In *Proceedings of LREC*.

David Chiang, Mona T Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic Dialects. In *Proceedings of EACL*.

Alina Maria Ciobanu and Liviu P. Dinu. 2016. A Computational Perspective on Romanian Dialects. In *Proceedings of LREC*.

Ryan Cotterell and Chris Callison-Burch. 2014. A Multi-dialect, Multi-genre Corpus of Informal Written Arabic. In *Proceedings LREC*.

Kareem Darwish, Hassan Sajjad, and Hamdy Mubarak. 2014. Verifiably Effective Arabic Dialect Identification. In *Proceedings of EMNLP*.

Heba Elfardy and Mona T Diab. 2013. Sentence Level Dialect Identification in Arabic. In *Proceedings of ACL*.

Cyril Goutte, Serge Léger, and Marine Carpuat. 2014. The NRC System for Discriminating Similar Languages. In *Proceedings of the VarDial Workshop*.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating Similar Languages: Evaluations and Explorations. In *Proceedings of LREC*.

Nizar Y Habash. 2010. Introduction to Arabic Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

Josef Kittler, Mohamad Hatef, Robert PW Duin, and Jiri Matas. 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239.

Ludmila I Kuncheva. 2004. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons.

Ludmila I Kuncheva. 2014. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, second edition.

Nikola Ljubešić, Nives Mikelic, and Damir Boras. 2007. Language Identification: How to Distinguish Similar Languages? In *Proceedings of the International Conference on Information Technology Interfaces*.

Marco Lui and Paul Cook. 2013. Classifying English Documents by National Dialect. In *Proceedings of ALTW*.

Shervin Malmasi and Mark Dras. 2015a. Automatic Language Identification for Persian and Dari Texts. In *Proceedings PACLING*.

Shervin Malmasi and Mark Dras. 2015b. Language Identification using Classifier Ensembles. In *Proceedings of the LT4VarDial workshop*.

Shervin Malmasi, Eshrag Refaee, and Mark Dras. 2015. Arabic Dialect Identification using a Parallel Multidialectal Corpus. In *Proceedings of PACLING*.

Shervin Malmasi, Mark Dras, and Marcos Zampieri. 2016a. LTG at SemEval-2016 Task 11: Complex Word Identification with Classifier Ensembles. In *Proceedings of SemEval*.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016b. Discriminating between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task. In *Proceedings of the VarDial Workshop*.

Gustavo Henrique Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex Word Identification. *Proceedings of SemEval*, pages 560–569.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic Identification of Arabic Language Varieties and Dialects in Social Media. In *Proceedings of the SocialNLP Workshop*.

Abdulhadi Shoufan and Sumaya Al-Ameri. 2015. Natural Language Processing for Dialectical Arabic: A Survey. In *Proceedings of the Arabic NLP Workshop*.

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 BUCC Workshop*.

Christoph Tillmann, Saab Mansour, and Yaser Al-Onaizan. 2014. Improved Sentence-Level Arabic Dialect Classification. In *Proceedings of the VarDial Workshop*.

Omar F Zaidan and Chris Callison-Burch. 2011. The Arabic Online Commentary Dataset: An Annotated Dataset of Informal Arabic with High Dialectal Content. In *Proceedings of ACL*.

Omar F Zaidan and Chris Callison-Burch. 2014. Arabic Dialect Identification. *Computational Linguistics*.

Marcos Zampieri and Binyam Gebrekidan Gebre. 2012. Automatic Identification of Language Varieties: The Case of Portuguese. In *Proceedings of KONVENS*.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A Report on the DSL Shared Task 2014. In *Proceedings of the VarDial Workshop*.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL Shared Task 2015. In *Proceedings of the LT4VarDial Workshop*.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic Dialects. In *Proceedings of NAACL-HLT*.