Discriminating Between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task

Shervin Malmasi[1,2], Marcos Zampieri[3], Nikola Ljubešić[4,5]
Preslav Nakov[7], Ahmed Ali[7], Jörg Tiedemann[6]

[1]Harvard Medical School, USA, [2]Macquarie University, Australia
[3]University of Cologne, Germany, [4]University of Zagreb, Croatia
[5]Jožef Stefan Institute, Slovenia, [6]University of Helsinki, Finland
[7]Qatar Computing Research Institute, HBKU, Qatar

Abstract

We present the results of the third edition of the Discriminating between Similar Languages (DSL) shared task, which was organized as part of the VarDial'2016 workshop at COLING'2016. The challenge offered two subtasks: subtask 1 focused on the identification of very similar languages and language varieties in newswire texts, whereas subtask 2 dealt with Arabic dialect identification in speech transcripts. A total of 37 teams registered to participate in the task, 24 teams submitted test results, and 20 teams also wrote system description papers. High-order character n-grams were the most successful feature, and the best classification approaches included traditional supervised learning methods such as SVM, logistic regression, and language models, while deep learning approaches did not perform very well.

1 Introduction

The Discriminating between Similar Languages (DSL) shared task on language identification was first organized in 2014. It provides an opportunity for researchers and developers to test language identification approaches for discriminating between similar languages, language varieties, and dialects. The task was organized by the workshop series on NLP for Similar Languages, Varieties and Dialects (VarDial), which was collocated in 2014 with COLING, in 2015 with RANLP, and in 2016 again with COLING.

In its third edition, the DSL shared task grew in size and scope featuring two subtasks and attracting a record number of participants. Below we present the task setup, the evaluation results, and a brief discussion about the features and learning methods that worked best. More detail about each particular system can be found in the corresponding system description paper, as cited in this report.

2 Related Work

Language and dialect identification have attracted a lot of research attention in recent years, covering a number of similar languages and language varieties such as South-Slavic languages (Ljubešić et al., 2007), English varieties (Lui and Cook, 2013), varieties of Mandarin in China, Taiwan and Singapore (Huang and Lee, 2008), Malay vs. Indonesian (Ranaivo-Malançon, 2006), Brazilian vs. European Portuguese (Zampieri and Gebre, 2012), and Persian vs. Dari (Malmasi and Dras, 2015a), to mention just a few. The interest in this aspect of language identification has motivated the organization of shared tasks such as the DSL challenge, which allowed researchers to compare various approaches using the same dataset.

Along with the interest in similar languages and language variety identification, we observed substantial interest in applying natural language processing (NLP) methods for the processing of dialectal Arabic with special interest in methods to discriminate between Arabic dialects. Shoufan and Al-Ameri (2015) presented a comprehensive survey on these methods including recent studies on Arabic dialect identification such as (Elfardy and Diab, 2014; Darwish et al., 2014; Zaidan and Callison-Burch, 2014;

Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects,
pages 1–14, Osaka, Japan, December 12 2016.

Tillmann et al., 2014; Malmasi and Dras, 2015a). Methods for Arabic dialect detection present significant overlap with methods proposed for similar language identification. For this reason, in the 2016 edition of the DSL challenge we offered a subtask on Arabic dialect identification.

Below, we discuss some related shared tasks including the first two editions of the DSL challenge.

2.1 Related Shared Tasks

Several shared tasks related to the DSL task have been organized in recent years. Two examples are the ALTW language identification shared task (Baldwin and Lui, 2010) on general-purpose language identification, and the DEFT 2010 shared task (Grouin et al., 2010), which focused on language variety identification of French texts with a temporal dimension. In the DEFT 2010 shared task, systems were asked to predict *when* and *where* texts were published. The DEFT 2010 shared task is most similar to our DSL task, but is limited to French language varieties, while our task is multilingual and includes several groups of similar languages and language varieties.

Language identification on Twitter and other platforms of user-generated content is a popular research direction (Ljubešić and Kranjčić, 2015). This interest has motivated the shared task on Language Identification in Code-Switched Data (Solorio et al., 2014), which focused on tweets containing a mix of two or more languages, and the TweetLID shared task (Zubiaga et al., 2014; Zubiaga et al., 2015), which targeted language identification of tweets focusing on English and on languages spoken on the Iberian peninsula, namely Basque, Catalan, Spanish, and Portuguese.

The most recent related shared task is the task on geolocation prediction in Twitter (Han et al., 2016).[1] The organizers of this task provided a large training set collected from one million users, and asked to predict the location of each user (user-level prediction) and of each tweet (*tweet*-level prediction).

2.2 Previous Editions of the DSL Task

For the first edition of the DSL task (Zampieri et al., 2014), we compiled v1.0 of the DSL corpus collection (DSLCC), which contained excerpts of newspaper texts written in thirteen languages divided into the following groups: Group A (Bosnian, Croatian, Serbian), Group B (Indonesian, Malay), Group C (Czech, Slovak), Group D (Brazilian Portuguese, European Portuguese), Group E (Peninsular Spanish, Argentinian Spanish), and Group F (American English, British English).[2]

Team	Closed	Open	System Description Paper
NRC-CNRC	**0.957**	-	(Goutte et al., 2014)
RAE	0.947	-	(Porta and Sancho, 2014)
UMich	0.932	0.859	(King et al., 2014)
UniMelb-NLP	0.918	**0.880**	(Lui et al., 2014)
QMUL	0.906	-	(Purver, 2014)
LIRA	0.766	-	-
UDE	0.681	-	-
CLCG	0.453	-	-
Total	**8**	**2**	**5**

Table 1: Results for the DSL 2014 shared task: accuracy.

Eight teams developed systems and submitted results for this first edition of the task. All eight teams participated in the closed track, which was limited to training on the DSL corpus only, and two teams took part in the open track, which also allowed using external resources; five teams submitted system description papers. The results are summarized in Table 1, where the best-performing submissions, in terms of accuracy, are shown in bold.[3]

The best score in the closed submission track was achieved by the *NRC-CNRC* team (Goutte et al., 2014), which used a two-step classification approach: they first predicted the language group, and then

[1] https://noisy-text.github.io/2016/geo-shared-task.html

[2] Group F was excluded from the official evaluation results due to a number of republications present in the dataset.

[3] For a comprehensive discussion of the first two editions of the DSL shared task, see (Goutte et al., 2016).

discriminated between the languages from this predicted language group. Members of this team also participated in 2015 under the name *NRC*. *UMich* (King et al., 2014) and *UniMelb-NLP* (Lui et al., 2014) were the only teams that compiled and used additional training resources and the only teams to make open submissions. However, their open submissions performed worse than their closed submissions: accuracy dropped from 93.2% to 85.9% for *UMich*, and from 91.8% to 88.0% for *UniMelb-NLP*.

For the 2015 edition of the task (Zampieri et al., 2015b), we created v2.0 of the DSLCC, which included the following languages and language varieties grouped by similarity: Bulgarian vs. Macedonian, Bosnian vs. Croatian vs. Serbian, Czech vs. Slovak, Malay vs. Indonesian, Brazilian vs. European Portuguese, Argentinian vs. Peninsular Spanish, and a group of various other languages,[4] which were included to emulate a more realistic language identification scenario. We had two test datasets. Test set A contained the original unmodified text excerpts, while in test set B we replaced the capitalized named entities by placeholders. The results for the participating systems in the 2015 edition of the DSL task are presented in Table 2; again, the best submissions are shown in bold. We can see that the 2015 edition of the task attracted more submissions compared to 2014.

Team	Closed A	Closed B	Open A	Open B	System Description Paper
BOBICEV	0.941	0.922	-	-	(Bobicev, 2015)
BRUNIBP	0.937	-	-	-	(Ács et al., 2015)
INRIA	0.839	-	-	-	-
MAC	**0.955**	**0.940**	-	-	(Malmasi and Dras, 2015b)
MMS*	0.952	0.928	-	-	(Zampieri et al., 2015a)
NLEL	0.640	0.628	0.918	0.896	(Fabra-Boluda et al., 2015)
NRC	0.952	0.930	**0.957**	**0.934**	(Goutte and Léger, 2015)
OSEVAL	-	-	0.762	0.753	-
PRHLT	0.927	0.908	-	-	(Franco-Salvador et al., 2015)
SUKI	0.947	0.930	-	-	(Jauhiainen et al., 2015)
Total	**9**	**7**	**3**	**3**	**8**

Table 2: Results for the DSL 2015 shared task: accuracy.

The best-performing system in the open submission track, that of *MAC*, used an ensemble of SVM classifiers and achieved 95.5% accuracy on test set A and 94.0% accuracy on test set B. Unlike in the 2014 edition, in which open submissions performed substantially worse than closed ones, this time this was not the case, e.g., for the *NRC* team. However, the additional resource they used was external only technically; in fact, it was the previous version of the DSL corpus.[5]

Moreover, the use of two test sets allowed us to evaluate the impact of named entities. In the 2014 edition of the task, we had noticed that names of people, places, and organizations could be quite helpful for discriminating texts from different geographical locations, e.g., Argentinian vs. Peninsular Spanish, and we were worried that this is what systems critically relied on, i.e., that they were focusing on country of origin rather than language variety prediction. However, the results for test set A vs. B in 2015 show that the influence of named entities was not as great as we feared, and that the participating systems were able to capture lexical and, in some cases syntactic, variation using n-gram models even when the original named entities were not present.

3 Task Setup

Here, we describe the setup of the 2016 DSL shared task: the subtasks, the tracks, and the data.

3.1 General Setup

This year, the DSL challenge included two subtasks:

[4]This group of languages included Catalan, Russian, Slovene, and Tagalog.

[5]The *NLEL* team reported a bug in their closed submission, which might explain their low performance in this track.

- **Subtask 1:** *Discriminating between Similar Languages and Language Varieties*
  For this subtask, we compiled a new version of the DSL corpus, which included for the first time French language varieties, namely Hexagonal French vs. Canadian French, and further excluded pairs of similar languages that proved to be very easy to discriminate between in previous editions (e.g., Czech vs. Slovak and Bulgarian vs. Macedonian).

- **Subtask 2:** *Arabic Dialect Identification*
  This subtask focused on discriminating Arabic dialects in speech transcripts. We used the dataset compiled by Ali et al. (2016), which contained Modern Standard Arabic (MSA) and four Arabic dialects: Egyptian (EGY), Gulf (GLF), Levantine (LAV), and North African (NOR).

As in previous editions of the DSL task, we allowed teams to use external data. We therefore divided each subtask in two tracks:

- **Closed:** using only the corpora provided by the organizers;

- **Open:** using any additional data.[6]

Participation this year increased substantially compared to previous years, as the statistics in Table 3 show. We believe that this is due to the addition of an Arabic subtask as well as to the out-of-domain tweet test sets for the English subtask.

Year	Venue	(Sub-)tasks	Subscriptions	Submissions	Papers
2014	VarDial at COLING	1	22	8	5
2015	LT4VarDial at RANLP	1	24	10	8
2016	VarDial at COLING	2	37	24	20

Table 3: The evolution of the DSL task from 2014 to 2016.

3.2 Data

In this section, we present the datasets we used this year. For subtask 1, we compiled v3.0 of DSLCC with a new language variety (French) as well as out-of-domain test sets with tweets, and for subtask 2, we use a corpus of Arabic transcribed speeches presented in (Ali et al., 2016).

3.2.1 Subtask 1 Data

We compiled a new version 3.0 of DSLCC, following the methodology we used in previous years (Tan et al., 2014).[7] The resulting corpus contains short newspaper texts written in twelve languages and language varieties. Table 4 shows the languages included in the DSL v3.0 grouped by similarity.

We provided participants with 20,000 instances per language variety divided into 18,000 instances for training and 2,000 for development. Most language groups included in v3.0 were also present in v1.0 and v2.0. We further added French from Canada and from France, as well as Mexican Spanish.[8]

We used three test sets for subtask 1: one in-domain (A), and two out-of-domain (B1 and B2). Test set A contained journalistic data including 1,000 instances per language sampled from the same distribution as for the DSLCC v3.0. It is also comparable to the test sets released in DSLCC v1.0 and v2.0.

We further created test sets B1 and B2 in order to evaluate the performance of the participating systems on out-of-domain data. Each of the two datasets included 100 Twitter users per language/variant, and a varying number of tweets per user. Note that these test sets cover only two groups of closely-related languages: South-Slavic (Bosnian, Croatian, Serbian) and Portuguese (Brazilian and European).

We used the TweetGeo (Ljubešić et al., 2016) and TweetCat (Ljubešić et al., 2014) tools for data collection. TweetGeo allows us to collect geo-encoded tweets over a specified perimeter via the Twitter

[6]For subtask 1, using previous versions of the DSL corpus also made a submission open.

[7]The dataset is available at `http://ttg.uni-saarland.de/resources/DSLCC`

[8]Mexican Spanish was already present for the unshared task in 2015, but now it is part of the main DSL shared task.

		Train and Dev.		Testing					
Language/Variety	**Class**	**Instances**	**Tokens**	**A**	**Tokens**	**B1**	**Tokens**	**B2**	**Tokens**
Bosnian	bs	20,000	743,732	1,000	37,630	100	209,884	100	170,481
Croatian	hr	20,000	874,555	1,000	42,703	100	179,354	100	119,837
Serbian	sr	20,000	813,076	1,000	41,153	100	181,185	100	124,469
Indonesian	id	20,000	831,647	1,000	42,192	—	—	—	—
Malay	my	20,000	618,532	1,000	31,162	—	—	—	—
Brazilian Portuguese	pt-BR	20,000	988,004	1,000	49288	100	151,749	100	19,567
European Portuguese	pt-PT	20,000	908,605	1,000	45173	100	134,139	100	13,145
Argentine Spanish	es-AR	20,000	999,425	1,000	50,135	—	—	—	—
Castilian Spanish	es-ES	20,000	1,080,523	1,000	53,731	—	—	—	—
Mexican Spanish	es-MX	20,000	751,718	1,000	47,176	—	—	—	—
Canadian French	fr-CA	20,000	772,467	1,000	38,602	—	—	—	—
Hexagonal French	fr-FR	20,000	963,867	1,000	48,129	—	—	—	—
Total		240,000	10,346,151	12,000	527,074	500	856,331	500	323,030

Table 4: DSLCC v3.0: the languages included in the corpus grouped by similarity. Note that a test example in test set A is an excerpt of text, whereas in test sets B1 and B2 it is a collection of multiple tweets by the same user (with 98.88 and 50.47 tweets per user on average for B1 and B2, respectively).

Stream API. We set up one perimeter over the South-Slavic speaking countries, another one over Portugal, and a final one over Brazil. We then collected data over a period of one month. Once ready, we filtered the users by number of tweets collected per user and by language the user predominantly used. Finally, we used the TweetCat tool to collect whole timelines for users matching the following criteria: the user has posted at least five tweets (otherwise language identification would be hard), and the language(s) given langid.py's prediction are (hr, sr, bs) for the first variant and (pt) for the second one.

We then proceeded to manual annotation. We had a single human annotator for each language/variety group. The annotation procedure was the following: the annotator read one tweet after the other, starting with the most recent tweet, and marking the tweet at which he made the decision about the language/variety used by the Twitter user. In the South-Slavic group, the average number of analyzed tweets per user was 70.5 for Bosnian, 51.5 for Croatian, and 49 for Serbian. In the Portuguese group, these were 6 for European Portuguese, and 8 for Brazilian Portuguese. While part of the difference between the two groups may be due to different criteria the two annotators used, the differences inside groups show important trends, e.g., that identifying Bosnian users requires on average 40% more tweets compared to identifying Serbian or Croatian ones.

Having the information about the number of tweets that were needed for a human decision enabled us to prepare the harder B2 test set in which only that minimum number of tweets was included. On the other hand, the B1 test set, being a proper superset of B2, contained much more tweets per user, and we had to cap the overall number of tweets in the dataset at 50,000 due to restrictions of the Twitter Developer Agreement.

It is important to stress that no filtering over the user timeline (such as removing tweets written in different languages or with no linguistic information) was performed, offering thereby a realistic setting.

3.2.2 Subtask 2 Data

For the Arabic subtask, we used transcribed speech in MSA and in four dialects (Ali et al., 2016): Egyptian (EGY), Gulf (GLF), Levantine (LAV), and North African (NOR). The data comes from a multi-dialectal speech corpus created from high-quality broadcast, debate and discussion programs from Al Jazeera, and as such contains a combination of spontaneous and scripted speech (Wray and Ali, 2015). We released 7,619 sentences for training and development, without a train/dev split;[9] a breakdown for each dialect is shown in Table 5. We further used 1,540 sentences for evaluation. We extracted text from ten hours of speech per dialect for training, and from two hours per dialect for testing.

[9] http://alt.qcri.org/resources/ArabicDialectIDCorpus/varDial_DSL_shared_task_2016_subtask2

Note that even though the origin of our data is speech, in our corpus we only used written transcripts. This makes the task hard as it may be difficult, or even impossible in certain contexts, to determine unambiguously the dialect of a written sentence if it contains graphemic cognates common across multiple dialects of colloquial and of Standard Arabic. This ambiguity is less pronounced in the presence of speech signal. Thus, we plan to make available acoustic features in future challenges.

		Training		Testing	
Dialect	**Dialect**	**Examples**	**Words**	**Examples**	**Words**
Egyptian	EGY	1,578	85K	315	13K
Gulf	GLF	1,672	65K	256	14K
Levantine	LAV	1,758	66K	344	14K
Modern Standard	MSA	999	49K	274	14K
North African	NOR	1,612	52K	351	12K
Total		7,619	317K	1,540	67K

Table 5: The Arabic training and testing data.

3.3 Evaluation

Regarding evaluation, in the previous editions of the DSL task, we used average accuracy as the main evaluation metric. This was because the DSL datasets were balanced with the same number of examples for each language variety. However, this is not true for this year's Arabic dataset, and thus we added macro-averaged F1-score, which is the official score this year.

Moreover, following common practice in other shared tasks, e.g., at WMT (Bojar et al., 2016), this year we carried out statistical significance tests using McNemar's test in order to investigate the variation of performance between the participating systems. Therefore, in all tables with results, we rank teams in groups taking statistical significance into account,[10] rather than using absolute performance only.

4 Results for Subtask 1: DSL Dataset

A total of 17 teams participated in the shared task. Table 6 shows statistics about the participating teams.

Team	**A (Closed)**	**A (Open)**	**B (Closed)**	**B (Open)**	**System Description Paper**
andre	✓				(Cianflone and Kosseim, 2016)
ASIREM	✓				(Adouane et al., 2016)
Citius_Ixa_Imaxin	✓	✓	✓	✓	(Gamallo et al., 2016)
eire	✓		✓		(Franco-Penya and Sanchez, 2016)
GW_LT3	✓		✓		(Zirikly et al., 2016)
HDSL	✓		✓		—
hltcoe	✓		✓		(McNamee, 2016)
mitsls	✓				(Belinkov and Glass, 2016)
nrc	✓	✓	✓	✓	(Goutte and Léger, 2016)
PITEOG	✓		✓	✓	(Herman et al., 2016)
ResIdent	✓		✓		(Bjerva, 2016)
SUKI	✓	✓	✓	✓	(Jauhiainen et al., 2016)
tubasfs	✓		✓		(Çöltekin and Rama, 2016)
UniBucNLP	✓		✓		(Ciobanu et al., 2016)
Uppsala	✓		✓		—
UPV_UA	✓		✓		—
XAC	✓		✓		(Barbaresi, 2016)
Total	**17**	**3**	**14**	**4**	**14**

Table 6: Teams participating in subtask 1 (here, we group test sets B1 and B2 under B).

[10] This means that systems not significantly different to the top system are also assigned rank 1, and so on.

4.1 Results on Test Set A

We received submissions by 17 teams for the closed training training condition. The results and a brief description of the algorithm and of the features used by each team are shown in Table 7. Note that the teams were allowed to submit up to three runs, and here we only show the results for the best run from each participating team. The best results in the closed condition were achieved by the *tubasfs* team with F1-score of 89.38% and by the *SUKI* team with F1-score of 88.77% (both ranked first, as they are not statistically different). A group of five teams scored between 88.14% and 88.70%, and they were all ranked second (as they were not statistically different).

Rank	Team	Run	Accuracy	F1	Approach
1	tubasfs	run1	**0.894**	**0.894**	SVM, char n-grams (1-7)
	SUKI	run1	0.888	0.888	Lang. models, word uni-, char n-grams (1-6)
2	GW_LT3	run3	0.887	0.887	Hierarchical log. regression, char/word n-grams
	nrc	run1	0.886	0.886	Two-stage probabilistic and SVM, char 6-grams
	UPV_UA	run1	0.883	0.884	String kernels and kernel discriminant analysis
	PITEOG	run3	0.883	0.883	Chunk-based language model
	andre	run1	0.885	0.881	Language models, char n-grams
3	XAC	run3	0.879	0.879	Unsupervised morphological model
	ASIREM	run1	0.878	0.878	SVM, char 4-grams
	hltcoe	run1	0.877	0.877	Prediction by partial matching, char 5-grams
4	UniBucNLP	run2	0.865	0.864	Hierarchical log. reg. w/ word 1/2-grams
5	HDSL	run1	0.853	0.852	SVM, word and char n-grams
	Citius_Ixa_Imaxin	run2	0.853	0.850	Naive Bayes, word unigrams
	ResIdent	run3	0.849	0.846	Deep neural net with byte embeddings
6	eire	run1	0.838	0.832	Naive Bayes, char bigrams
	mitsls	run3	0.830	0.830	Character-level convolutional neural network
7	Uppsala	run2	0.825	0.824	Word-level convolutional neural network

Table 7: Results for subtask 1, test set A, *closed* training condition.

Rank	Team	Run	Accuracy	F1	Approach
1	nrc	run1	**0.890**	**0.889**	Two-stage probabilistic and SVM, char 6-grams
	SUKI	run1	0.884	0.884	Lang. models, word uni-, char n-grams (1-7)
2	Citius_Ixa_Imaxin	run2	0.871	0.869	Naive Bayes, word unigrams

Table 8: Results for subtask 1, test set A, *open* training condition.

The open training track for test set A attracted only three teams as shown in Table 8. For the first two teams, the difference compared to their closed submission is marginal: *nrc* gained less than half a point absolute in terms of accuracy and F1, while *SUKI* lost about the same. However, the third team, *Citius_Ixa_Imaxin*, managed to gain about two points absolute in both measures.

Overall, we observe that the teams used a wide variety of algorithms and features, which are summarized in the results tables. They are also described in more detail in the corresponding system description papers. Note that some teams, such as *ResIdent* and *Uppsala*, used neural network-based approaches, but their results were not competitive to those that used simpler, standard classifiers such as SVM and logistic regression.

4.2 Results on Test Sets B1 and B2

The results of the participating teams on test set B1 (out-of-domain, tweets) for the closed training condition are shown in Table 9. Once again, we group the submissions based on statistical significance. Three teams shared the first place, namely *GW_LT3*, *nrc*, and *UniBucNLP*, with an F1-score ranging from 89.69% to 91.94%.

Rank	Team	Run	Accuracy	F1	Approach
1	GW_LT3	run1	**0.920**	**0.919**	Log. reg. with char/word n-grams
	nrc	run1	0.914	0.913	Two-stage probabilistic and SVM, char 6-grams
	UniBucNLP	run1	0.898	0.897	Log. reg. w/ word 1/2-grams
2	UPV_UA	run2	0.888	0.886	String kernels and kernel discriminant analysis
	tubasfs	run1	0.862	0.860	SVM, char n-grams (1-7)
3	eire	run1	0.806	0.793	Naive Bayes, char bigrams
	PITEOG	run1	0.800	0.793	Expectation maximization, word unigrams
4	Citius_Ixa_Imaxin	run1	0.708	0.713	Dictionary-based ranking method
	ResIdent	run3	0.688	0.687	Deep neural net with byte embeddings
	HDSL	run1	0.698	0.686	SVM, word and char n-grams
	Uppsala	run2	0.682	0.685	Word-level convolutional neural network
	SUKI	run3	0.688	0.672	Lang. models, word uni-, char n-grams (1-8)
5	XAC	run2	0.618	0.594	Unsupervised morphological model
6	hltcoe	run1	0.530	0.510	Prediction by partial matching, char 5-grams

Table 9: Results for subtask 1, test set B1, *closed* training condition.

Rank	Team	Run	Accuracy	F1	Approach
1	nrc	run1	**0.948**	**0.948**	Two-stage probabilistic and SVM, char 6-grams
2	SUKI	run3	0.822	0.815	Lang. models, word uni-, char n-grams (1-8)
	PITEOG	run1	0.800	0.815	Expectation maximization, word unigrams
4	Citius_Ixa_Imaxin	run1	0.664	0.634	Dictionary-based ranking method

Table 10: Results for subtask 1, test set B1, *open* training condition.

Rank	Team	Run	Accuracy	F1	Approach
1	GW_LT3	run1	**0.878**	**0.877**	Log. reg. with char/word n-grams
	nrc	run1	**0.878**	**0.877**	Two-stage probabilistic and SVM, char 6-grams
	UPV_UA	run2	0.858	0.857	String kernels and kernel discriminant analysis
2	UniBucNLP	run2	0.838	0.838	Hierarchical log. reg. w/ word 1/2-grams
	tubasfs	run1	0.822	0.818	SVM, char n-grams (1-7)
3	PITEOG	run1	0.760	0.757	Expectation maximization, word unigrams
	eire	run1	0.740	0.727	Naive Bayes, char bigrams
4	Citius_Ixa_Imaxin	run1	0.686	0.698	Dictionary-based ranking method
	ResIdent	run2	0.698	0.694	Deep neural net with byte embeddings
	Uppsala	run2	0.672	0.675	Word-level convolutional neural network
	HDSL	run1	0.640	0.626	SVM, word and char n-grams
	SUKI	run1	0.642	0.623	Lang. models, word uni-, char n-grams (1-6)
5	XAC	run2	0.576	0.552	Unsupervised morphological model
	hltcoe	run2	0.554	0.513	Prediction by partial matching, char 5-grams

Table 11: Results for subtask 1, test set B2, *closed* training condition.

Rank	Team	Run	Accuracy	F1	Approach
1	nrc	run1	**0.900**	**0.900**	Two-stage probabilistic and SVM, char 6-grams
2	SUKI	run2	0.796	0.791	Lang. models, word uni-, char n-grams (1-8)
3	PITEOG	run1	0.728	0.759	Expectation maximization, word unigrams
	Citius_Ixa_Imaxin	run1	0.692	0.695	Dictionary-based ranking method

Table 12: Results for subtask 1, test set B2, *open* training condition.

Note that the higher results obtained on test set B1 compared to test set A are somewhat misleading: test set B1 is out-of-domain and is thus generally harder, but it also involves less languages (five for test set B1 as opposed to twelve for test set A), which makes it ultimately much easier.

In Table 10 we present the results for the four teams that participated under the open training condition for test set B1, i.e., using external data. We can see that the *nrc* team performed best with an F1 score of 94.80%. This result is a few percentage points better than the 91.34% F1-score obtained by *nrc* in the closed training condition, which indicates that the use of additional training data was indeed helpful. This is an expected outcome as no suitable training data has been provided for test sets B1 and B2, which contain tweets, and are out of domain compared to the training data (newspaper texts).

Table 11 shows the results on test set B2 under the closed training condition. As expected, this test set turned out to be more challenging than test set B1, and this was the case for almost all teams. Moreover, we can see that there was some minor variation in the ranks of teams on B1 and on B2 (closed training condition), e.g., the *UniBucNLP* team was ranked among the first on B1, but for B2 it switched places with the *UPV UA* team.

Finally, Table 12 presents the results on test set B2 in the open training condition. Once again, the results of *nrc* were higher here than in the closed training condition.

4.3 Open Training Data Sources

Collecting additional training data is a time-consuming process. Therefore, in line with our expectations given our past experience in the previous editions of the DSL task, we received far fewer entries in the open training condition for both subtasks.

For subtask 1, a total of four teams used additional training data across the three test sets. According to the system description papers, the data was compiled from the following sources:

- *Citius_Ixa_Imaxin* augmented the training data with the corpus released in the second edition of the DSL task in 2015.
- *nrc* augmented the provided training data with the corpora from the two previous DSL shared tasks (DSLCC v1.0 and DSLCC v2.1), plus additional text crawled from the web site of the newspaper *La Presse* from Quebec.
- *PITEOG* used their own custom web-based corpus, with no further details provided.
- *SUKI* created an additional dataset using web pages in the Common Crawl corpus.

5 Results for Subtask 2: Arabic Dialect Identification

The eighteen teams that participated in subtask 2 along with the reference to their system description papers are shown in Table 13.[11]

5.1 Results on Subtask C

The results obtained by the teams that participated in the closed training condition are shown in Table 14. The best results were obtained by *MAZA*, *UnibucKernel*, *QCRI*, and *ASIREM*, which achieved an F1-score ranging between 49.46% and 51.32%, and thus shared the first place. The *MAZA* team proposed an approach based on SVM ensembles, which was also ranked first in the 2015 edition of the DSL task (Malmasi and Dras, 2015b), which confirms that SVM ensembles are a suitable method for this task. The *UnibucKernel* team approached the task using string kernels, which were previously proposed for native language identification (Ionescu et al., 2016).

Table 15 shows the results obtained by the three teams that participated in subtask 2 under the open training condition. They showed very different performance (statistically different), and saw very different outcomes when using external training data.

[11] We acknowledge that team *MAZA* included two DSL shared task organizers. Yet, the team had no unfair advantage, and competed under the exactly same conditions as the other participants.

Team	C (Closed)	C (Open)	System Description Paper
AHAQST	✓		(Hanani et al., 2016)
ALL	✓		(Alshutayri et al., 2016)
ASIREM	✓	✓	(Adouane et al., 2016)
cgli	✓		(Guggilla, 2016)
Citius_Ixa_Imaxin	✓		(Gamallo et al., 2016)
eire	✓		(Franco-Penya and Sanchez, 2016)
GW_LT3	✓	✓	(Zirikly et al., 2016)
HDSL	✓		—
hltcoe	✓		(McNamee, 2016)
MAZA	✓		(Malmasi and Zampieri, 2016)
mitsls	✓		(Belinkov and Glass, 2016)
PITEOG	✓		(Herman et al., 2016)
QCRI	✓	✓	(Eldesouki et al., 2016)
SUKI	✓		(Jauhiainen et al., 2016)
tubasfs	✓		(Çöltekin and Rama, 2016)
UCREL	✓		—
UnibucKernel	✓		(Ionescu and Popescu, 2016)
UniBucNLP	✓		(Ciobanu et al., 2016)
Total	**18**	**3**	**15**

Table 13: The teams that participated in subtask 2 (Arabic).

Rank	Team	Run	Accuracy	F1	Approach
1	MAZA	run3	0.512	**0.513**	Ensemble, word/char n-grams
	UnibucKernel	run3	0.509	0.513	Multiple string kernels
	QCRI	run1	**0.514**	0.511	SVM, word/char n-grams
	ASIREM	run1	0.497	0.495	SVM, char 5/6-grams
2	GW_LT3	run3	0.490	0.492	Ensemble, word/char n-grams
	mitsls	run3	0.485	0.483	Character-level convolutional neural network
	SUKI	run1	0.488	0.482	Language models, char n-grams (1-8)
	UniBucNLP	run3	0.475	0.474	SVM w/ string kernels (char 2-7 grams)
	tubasfs	run1	0.475	0.473	SVM, char n-grams (1-7)
3	HDSL	run1	0.458	0.459	SVM, word and char n-grams
	PITEOG	run2	0.461	0.452	Expectation maximization, word unigrams
4	ALL	run1	0.429	0.435	SVM, char trigrams
	cgli	run3	0.438	0.433	Convolutional neural network (CNN)
	AHAQST	run1	0.428	0.426	SVM, char trigrams
	hltcoe	run1	0.412	0.413	Prediction by partial matching, char 4-grams
5	Citius_Ixa_Imaxin	run1	0.387	0.382	Dictionary-based ranking method
5	eire	run1	0.358	0.346	Naive Bayes, char bigrams
6	UCREL	run2	0.261	0.244	Decision tree (J48), word frequencies

Table 14: Results for subtask 2 (Arabic), *closed* training condition.

Rank	Team	Run	Accuracy	F1	Approach
1	ASIREM	run3	**0.532**	**0.527**	SVM, char 5/6-grams
2	GW_LT3	run3	0.491	0.493	Ensemble, word/char n-grams
3	QCRI	run1	0.379	0.352	SVM, word/char n-grams

Table 15: Results for subtask 2 (Arabic), *open* training condition.

The best-performing system proposed by the *ASIREM* team achieved higher results in the open vs. the closed training condition (52.74% vs. 49.46% F1-score); the second-best system by the *GW_LT3* team performed very similarly in the two conditions (an F1-score of 49.29% for open and 49.22% for closed training); and the third team, *QCRI*, actually performed much better in the closed training condition than in the open one (51.12% vs. 35.20% F1-score). This variation can be explained by looking at the additional training data these teams used, which we will do in the next subsection.

5.2 Open Training Data Sources

The three teams who participated in the open training condition used the following sources:

- *ASIREM* used 18,000 documents (609,316 words) collected manually by native speakers from social media. This yielded results that outperformed the best system in the closed training track, thus demonstrating that out-of-domain training data can be quite useful for this task.
- The *GW_LT3* team made use of dialectal dictionaries and data they collected from Twitter, which also worked quite well.
- The *QCRI* team used a multi-dialect, multi-genre corpus of informal written Arabic (Zaidan and Callison-Burch, 2011).

6 Approaches and Trends

6.1 Features

Almost all teams relied on standard word and character n-grams. Key trends here were that character n-grams outperformed their word-based counterparts, and that higher-order n-grams (5-, 6- and 7-grams) did very well. In fact, the top teams in all categories made use of high-order n-grams. The two teams that were ranked first in test set A used only character n-grams of order 1–7, which demonstrates that combining the n-grams of different orders can be useful.

6.2 Machine Learning Approaches: Traditional vs. Deep Learning Methods

When analyzing the results, we observed several trends about how machine learning approaches were used. For example, we found that traditional supervised learning approaches, particularly SVM and logistic regression, performed very well. In fact, the winner of each category used one of these approaches. This is not surprising given that these methods are suitable for tasks with large numbers of features. Complex learning approaches, such as ensemble methods or hierarchical classifiers, also performed well. Many of the winning runs or those in the top-3 for each category used such an approach.

In contrast, numerous teams attempted to use new deep learning-based approaches, with most of them performing poorly compared to traditional classifiers. One exception is the character-level CNN used by the *mitsls* team, which ranked in sixth place for test set C. Several teams submitted runs using both simple classifiers and deep learning methods, with most noting that the simple methods proved difficult to beat even when comparing against very sophisticated neural network architectures. Others noted the memory requirements and long training times, which made the use of deep learning methods difficult. For example, one team mentioned that their model needed ten days to train.

7 Conclusion

The 2016 DSL shared task was once again a very fruitful experience for both the organizers and the participants. The record number of 37 subscriptions and 24 submissions confirms the interest of the community in discriminating between dialects and similar languages.

This year, we split the task into two subtasks: one on similar languages and varieties and one on Arabic dialect identification. For subtask 1, we provided an in-domain test set (A) compiled from news corpora and an out-of-domain test sets (B1 and B2) collected from social media; the latter case was more challenging. The new subtask on Arabic dialects and the new datasets we released brought even more attention to the DSL task, which ultimately resulted in a record number of submissions.

We are delighted to see many teams developing systems and testing approaches in both subtasks. We observed that more teams used deep learning in comparison to previous editions of the DSL task. Yet, the best results were obtained by simpler machine learning methods such as SVM and logistic regression.

Acknowledgments

We would like to thank all participants in the DSL shared task for their valuable suggestions and comments. We further thank the VarDial Program Committee for thoroughly reviewing the system papers and for their feedback on this report.

References

Judit Ács, László Grad-Gyenge, and Thiago Bruno Rodrigues de Rezende Oliveira. 2015. A Two-level Classifier for Discriminating Similar Languages. In *Proceedings of the LT4VarDial Workshop*.

Wafia Adouane, Nasredine Semmar, Richard Johansson, and Victoria Bobicev. 2016. ASIREM Participation at the Discriminating Similar Languages Shared Task 2016. In *Proceedings of the VarDial Workshop*.

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*.

Areej Alshutayri, Eric Atwell, Abdulrahman Alosaimy, James Dickins, Michale Ingleby, and Janet Watson. 2016. Arabic Language WEKA-Based Dialect Classifier for Arabic Automatic Speech Recognition Transcripts. In *Proceedings of the VarDial Workshop*.

Timothy Baldwin and Marco Lui. 2010. Multilingual Language Identification: ALTW 2010 Shared Task Data. In *Proceedings of ALTA*.

Adrien Barbaresi. 2016. An Unsupervised Morphological Criterion for Discriminating Similar Languages. In *Proceedings of the VarDial Workshop*.

Yonatan Belinkov and James Glass. 2016. A Character-level Convolutional Neural Network for Distinguishing Similar Languages and Dialects. In *Proceedings of the VarDial Workshop*.

Johannes Bjerva. 2016. Byte-based Language Identification with Deep Convolutional Networks. In *Proceedings of the VarDial Workshop*.

Victoria Bobicev. 2015. Discriminating between Similar Languages Using PPM. In *Proceedings of the LT4VarDial Workshop*.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of WMT*.

Çağrı Çöltekin and Taraka Rama. 2016. Discriminating Similar Languages with Linear SVMs and Neural Networks. In *Proceedings of the VarDial Workshop*.

Andre Cianflone and Leila Kosseim. 2016. N-gram and Neural Language Models for Discriminating Similar Languages. In *Proceedings of the VarDial Workshop*.

Alina Maria Ciobanu, Sergiu Nisioi, and Liviu P. Dinu. 2016. Vanilla Classifiers for Distinguishing between Similar Languages. In *Proceedings of the VarDial Workshop*.

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

Mohamed Eldesouki, Fahim Dalvi, Hassan Sajjad, and Kareem Darwish. 2016. QCRI @ DSL 2016: Spoken Arabic Dialect Identification Using Textual Features. In *Proceedings of the VarDial Workshop*.

Heba Elfardy and Mona Diab. 2014. Sentence Level Dialect Identification in Arabic. In *Proceedings of ACL*.

Raül Fabra-Boluda, Francisco Rangel, and Paolo Rosso. 2015. NLEL UPV Autoritas participation at Discrimination between Similar Languages (DSL) 2015 shared task. In *Proceedings of the LT4VarDial Workshop*.

Hector-Hugo Franco-Penya and Liliana Mamani Sanchez. 2016. Tuning Bayes Baseline for Dialect Detection. In *Proceedings of the VarDial Workshop*.

Marc Franco-Salvador, Paolo Rosso, and Francisco Rangel. 2015. Distributed Representations of Words and Documents for Discriminating Similar Languages. In *Proceedings of the LT4VarDial Workshop*.

Pablo Gamallo, Iñaki Alegria, and José Ramom Pichel. 2016. Comparing two Basic Methods for Discriminating Between Similar Languages and Varieties. In *Proceedings of the VarDial Workshop*.

Cyril Goutte and Serge Léger. 2015. Experiments in Discriminating Similar Languages. In *Proceedings of the LT4VarDial Workshop*.

Cyril Goutte and Serge Léger. 2016. Advances in Ngram-based Discrimination of Similar Languages. In *Proceedings of the VarDial Workshop*.

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*.

Cyril Grouin, Dominic Forest, Lyne Da Sylva, Patrick Paroubek, and Pierre Zweigenbaum. 2010. Présentation et Résultats du Défi Fouille de Texte DEFT2010 Où et Quand un Article de Presse a-t-il Été Écrit? In *Proceedings of DEFT*.

Chinnappa Guggilla. 2016. Discrimination between Similar Languages, Varieties and Dialects using CNN and LSTM-based Deep Neural Networks. In *Proceedings of the VarDial Workshop*.

Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. 2016. Twitter Geolocation Prediction Shared Task of the 2016 Workshop on Noisy User-generated Text. In *Proceedings of the W-NUT Workshop*.

Abualsoud Hanani, Aziz Qaroush, and Stephen Taylor. 2016. Classifying ASR Transcriptions According to Arabic Dialect. In *Proceedings of the VarDial Workshop*.

Ondřej Herman, Vit Suchomel, Vít Baisa, and Pavel Pavel Rychlý. 2016. DSL Shared task 2016: Perfect Is The Enemy of Good Language Discrimination Through Expectation–Maximization and Chunk-based Language Model. In *Proceedings of the VarDial Workshop*.

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*.

Radu Tudor Ionescu and Marius Popescu. 2016. UnibucKernel: An Approach for Arabic Dialect Identification based on Multiple String Kernels. In *Proceedings of the VarDial Workshop*.

Radu Tudor Ionescu, Marius Popescu, and Aoife Cahill. 2016. String Kernels for Native Language Identification: Insights from Behind the Curtains. *Computational Linguistics*, 43(3):491–525.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2015. Discriminating Similar Languages with Token-based Backoff. In *Proceedings of the LT4VarDial Workshop*.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2016. HeLI, a Word-Based Backoff Method for Language Identification. In *Proceedings of the VarDial Workshop*.

Ben King, Dragomir Radev, and Steven Abney. 2014. Experiments in Sentence Language Identification with Groups of Similar Languages. In *Proceedings of the VarDial Workshop*.

Nikola Ljubešić and Denis Kranjčić. 2015. Discriminating Between Closely Related Languages on Twitter. *Informatica*, 39(1).

Nikola Ljubešić, Nives Mikelic, and Damir Boras. 2007. Language identification: How to distinguish similar languages? In *Proceedings of ITI*.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2014. TweetCaT: a Tool for Building Twitter Corpora of Smaller Languages. In *Proceedings of LREC*.

Nikola Ljubešić, Tanja Samardžić, and Curdin Derungs. 2016. TweetGeo – A Tool for Collecting, Processing and Analysing Geo-encoded Data. In *Proceedings of COLING*.

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

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 VarDial*.

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

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

Shervin Malmasi and Marcos Zampieri. 2016. Arabic Dialect Identification in Speech Transcripts. In *Proceedings of the VarDial Workshop*.

Paul McNamee. 2016. Language and Dialect Discrimination Using Compression-Inspired Language Models. In *Proceedings of the VarDial Workshop*.

Jordi Porta and José-Luis Sancho. 2014. Using Maximum Entropy Models to Discriminate between Similar Languages and Varieties. In *Proceedings of the VarDial Workshop*.

Matthew Purver. 2014. A Simple Baseline for Discriminating Similar Language. In *Proceedings of VarDial Workshop*.

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

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

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the First Shared Task on Language Identification in Code-Switched Data. In *Proceedings of the CodeSwitch Workshop*.

Liling Tan, Marcos Zampieri, Nikola Ljubešić, 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*.

Samantha Wray and Ahmed Ali. 2015. Crowdsource a little to label a lot: labeling a speech corpus of dialectal Arabic. In *Proceedings of Interspeech*.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of ACL-HLT*.

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 *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, Binyam Gebrekidan Gebre, Hernani Costa, and Josef van Genabith. 2015a. Comparing Approaches to the Identification of Similar Languages. In *Proceedings of the VarDial Workshop*.

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

Ayah Zirikly, Bart Desmet, and Mona Diab. 2016. The GW/LT3 VarDial 2016 Shared Task System for Dialects and Similar Languages Detection. In *Proceedings of the VarDial Workshop*.

Arkaitz Zubiaga, Inaki San Vicente, Pablo Gamallo, José Ramom Pichel, Inaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Vıctor Fresno. 2014. Overview of TweetLID: Tweet Language Identification at SEPLN 2014. In *Proceedings of the TweetLID Workshop*.

Arkaitz Zubiaga, Iñaki San Vicente, Pablo Gamallo, José Ramom Pichel, Iñaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno. 2015. TweetLID: A Benchmark for Tweet Language Identification. *Language Resources and Evaluation*, pages 1–38.