A Report on the VarDial Evaluation Campaign 2020

Mihaela Găman[1], Dirk Hovy[2], Radu Tudor Ionescu[1], Heidi Jauhiainen[3]
Tommi Jauhiainen[3], Krister Lindén[3], Nikola Ljubešić[4,5], Niko Partanen[3]
Christoph Purschke[6], Yves Scherrer[3], Marcos Zampieri[7]

[1]University of Bucharest, [2]Bocconi University, [3]University of Helsinki,
[4]Jožef Stefan Institute, [5]University of Ljubljana, [6]University of Luxembourg,
[7]Rochester Institute of Technology

vardialworkshop@gmail.com

Abstract

This paper presents the results of the VarDial Evaluation Campaign 2020 organized as part of the seventh workshop on Natural Language Processing (NLP) for Similar Languages, Varieties and Dialects (VarDial), co-located with COLING 2020. The campaign included three shared tasks each focusing on a different challenge of language and dialect identification: Romanian Dialect Identification (RDI), Social Media Variety Geolocation (SMG), and Uralic Language Identification (ULI). The campaign attracted 30 teams who enrolled to participate in one or multiple shared tasks and 14 of them submitted runs across the three shared tasks. Finally, 11 papers describing participating systems are published in the VarDial proceedings and referred to in this report.

1 Introduction

The VarDial Evaluation Campaign 2020[1] is the most recent iteration of a series of evaluation campaigns featuring multiple shared tasks organized together with the Workshop on Natural Language Processing (NLP) for Similar Languages, Varieties and Dialects (VarDial). It follows three editions organized in 2017 (Zampieri et al., 2017) featuring four shared tasks and in 2018 (Zampieri et al., 2018) and 2019 (Zampieri et al., 2019) featuring five shared tasks.

Co-located with international NLP conferences such as COLING, EACL, and NAACL, VarDial is a forum for researchers interested in diatopic language variation from a computational perspective. Since its first edition in 2014, VarDial hosted shared tasks on various topics such as morphosyntactic tagging, cross-lingual dependency parsing, and language and dialect identification. Most shared tasks organized at VarDial have addressed dialect and language identification on newspaper texts, social media posts, speech transcriptions, and many other genres and domains (Malmasi et al., 2016; Goutte et al., 2016). A large number of languages and dialects from different families have been included in the VarDial shared tasks: national language varieties of Chinese, English, French, Spanish, and Portuguese, pairs or groups of similar languages such as Bosnian, Croatian, and Serbian and Malay and Indonesian, and dialects of languages such as Arabic and German. Some of the datasets made available in these tasks, such as the ArchiMob for Swiss German dialects and the multilingual DSLCC, have been used outside these competitions evidencing the interest of the NLP community in the topic (Tan et al., 2014; Samardžić et al., 2016; Kumar et al., 2018).[2]

In this paper, we present the results and the main findings of the VarDial Evaluation Campaign 2020. Three tasks addressing different aspects of language and dialect identification have been organized this year. The Romanian Dialect Identification (RDI) shared task is described in Section 4, the Social Media Variety Geolocation (SMG) task is presented in Section 5, and finally the Uralic Language Identification (ULI) shared task is described in Section 6. We include references to the 11 system description papers written by the participants of the campaign in Table 1.

[1]https://sites.google.com/view/vardial2020/evaluation-campaign

[2]For recent surveys on these topics see Zampieri et al. (2020) and Jauhiainen et al. (2019c).

Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 1–14
Barcelona, Spain (Online), December 13, 2020

2 Shared Tasks at VarDial 2020

Romanian Dialect Identification (RDI): In the Romanian Dialect Identification (RDI) shared task, we provided participants with the MOROCO data set (Butnaru and Ionescu, 2019) for training, which contains Moldavian (MD) and Romanian (RO) samples of text collected from the news domain. The task was a binary classification by dialect, in which a classification model is required to discriminate between the Moldavian (MD) and the Romanian (RO) dialects. The task was closed, therefore, participants are not allowed to use external data to train their models. The test set contained newly collected text samples from a different domain, not previously included in MOROCO, resulting in a cross-domain dialect identification task.

Social Media Variety Geolocation (SMG): In contrast to most past and present VarDial tasks, the SMG task is framed as a geolocation task: given a text, the participants have to predict its geographic location in terms of latitude/longitude coordinates. This setup addresses the common issue that defining a set of discrete labels is not trivial for many language areas where there is a continuum between varieties rather than clear-cut borders. The SMG task is split into three subtasks covering different language areas: the **BCMS** subtask is focused on geolocated tweets published in the area of Croatia, Bosnia and Herzegovina, Montenegro and Serbia in the HBS macro-language (Ljubešić et al., 2016); the **DE-AT** subtask focuses on conversations from the microblogging platform Jodel initiated in Germany and Austria, which are written in standard German but commonly contain regional and dialectal forms; the **CH** subtask is based on Jodel conversations initiated in Switzerland, which were found to be held majoritarily in Swiss German dialects (Hovy and Purschke, 2018). All three subtasks used the same data format and evaluation methodology. Both constrained and unconstrained submissions were allowed, but only one participating team made use of the latter.

Uralic Language Identification (ULI): This shared task focused on discriminating between endangered languages of the Uralic group. In addition to 29 Uralic minority languages, the shared task also featured 149 non-relevant languages. For training, we provided texts from the Wanca 2016 corpora (Jauhiainen et al., 2019a) for the relevant languages while the texts for the non-relevant languages came from the Leipzig corpora collection (Goldhahn et al., 2012). The test set for the relevant languages included sentences from the forthcoming Wanca 2017 corpora (Jauhiainen et al., 2020b) that were not present in the Wanca 2016 corpora. The sentences for the non-relevant languages were from the Leipzig corpora collection. The ULI shared task was divided into three separate tracks using the same training and test data. The difference between the tracks was based on how the submissions were scored: track 1 focused on macro-averaged F-score for the 29 relevant languages, track 2 on micro-averaged F-score for the relevant languages, and track 3 on macro-averaged F-score for all 178 languages. All the tracks were closed, so no other data or models were to be used for training in addition to the pre-defined training sets.

3 Participating Teams

A total of 30 teams enrolled to participate in this year's VarDial evaluation campaign and 14 of them submitted results to one or more shared tasks. In Table 1, we list the teams that participated in the shared tasks, including references to the 11 system description papers written by the participants. We include detailed information about these submissions in each respective task section of this report.

The RDI task attracted 8 teams followed by the SMG task with 7 teams who submitted runs to one or more of its three language tracks: BCMS, DE-AT, and CH. This is a similar number of teams that have participated in the most popular shared tasks from the VarDial evaluation campaign 2019. Only the NRC team submitted results to the ULI shared task, which is rather unusual, as tasks in past VarDial evaluation campaign have all received a very good number of submissions. It should be noted that the 2020 campaign run from April 20 to July 30 during the early stages of the COVID-19 pandemic. Lock downs and restrictive measures in many countries during this period have impacted universities and research centers worldwide causing significant disruption. We believe that this situation is very likely to have discouraged more teams to participate in this year's evaluation campaign.

Team	RDI	SMG	ULI	System Description Paper
Akanksha	✓			
Anumiţi	✓			(Popa and Ștefănescu, 2020)
CUBoulder-UBC		✓		*
Phlyers	✓			(Ceolin and Zhang, 2020)
HeLju		✓		(Scherrer and Ljubešić, 2020)
NRC			✓	(Bernier-Colborne and Goutte, 2020)
Piyush Mishra		✓		(Mishra, 2020)
SUKI	✓	✓		(Jauhiainen et al., 2020a)
The Linguistadors	✓	✓		
Tübingen	✓			(Çöltekin, 2020)
UAIC	✓			(Rebeja and Cristea, 2020)
UnibucKernel		✓		(Găman and Ionescu, 2020a)
UPB	✓			(Zaharia et al., 2020)
ZHAW-InIT		✓		(Benites et al., 2020)
Total	**8**	**7**	**1**	**11**

Table 1: The teams that participated in the VarDial Evaluation Campaign 2020 along with their system description papers. *The system description paper by team CUBoulder-UBC does not appear in the VarDial workshop proceedings. CUBoulder-UBC reused a system described in Hulden et al. (2015).

4 Romanian Dialect Identification RDI

4.1 Dataset

The training data is composed of news articles from the Moldavian and Romanian Dialectal Corpus (MOROCO)[3] (Butnaru and Ionescu, 2019). MOROCO was collected from the top five news websites from Romania and the Republic of Moldova, using each country's web domain (*.ro* or *.md*) to automatically separate the news articles by dialect.

The test data consists of short text samples from MOROCO-Tweets[4] (Găman and Ionescu, 2020b). The tweets are collected from Romania and the Republic of Moldova, the labels being assigned based on the geographical location of tweets.

Dialect	Training Set Size	Development Set Size		Test Set Size
		News Articles	Tweets	
Romanian	18,161	3,205	102	2,523
Moldavian	15,403	2,718	113	2,499
Total	**33,564**	**5,923**	**215**	**5,022**

Table 2: Number of text samples in the training, the development and the test sets considered for the RDI shared task.

The chosen training and test corpora allowed us to evaluate participants on a challenging cross-genre binary dialect identification task: Romanian (RO) versus Moldavian (MD). However, participants were provided with development data comprising both news articles and tweets. The number of samples in the training, the development and the test sets are listed in Table 2. All text samples were automatically pre-processed to replace each named entity with the token NE.

[3]https://github.com/butnaruandrei/MOROCO

[4]https://github.com/raduionescu/MOROCO-Tweets

4.2 Participants and Approaches

Akanksha. The Akanksha team fine-tuned a reformer model (Kitaev et al., 2019) on the provided data, considering character-level and phrase-level tokens. Then, a binary classifier is trained on top of the fine-tuned reformer model. The team submitted only one run.

Anumiţi. The Anumiţi team (Popa and Ștefănescu, 2020) submitted three runs using fine-tuned Romanian BERT models (Dumitrescu et al., 2020). They started from BERT models that are pre-trained on Romanian corpora. For the first two runs, the team submitted individual models, the first one being a cased BERT model and the second one being an uncased BERT model, respectively. For the third run, the team proposed an SVM ensemble of five different transformer-based models, some being multilingual and others being specifically trained on Romanian corpora.

Phlyers. All the submissions made by the Phlyers (Ceolin and Zhang, 2020) are based on Naïve Bayes models applied on character n-grams. Before applying the models, the team preprocessed the text samples by removing numbers, punctuation and common Twitter tags such as "LIVE", "FOTO" and "VIDEO", as well as the NE tag (which was used to replace named entities). For the first run, the Phlyers tuned the model on the news development set, obtaining optimal results with $\alpha = 10^{-4}$ and n-grams in the range 5-8 that occur less than 1000 times. For the second and third runs, the Phlyers tuned the model on the tweets development set. The best model uses $\alpha = 10^{-3}$ and n-grams in the range 6-8 that occur less than 250 times, while the second best model uses $\alpha = 10^{-3}$ and n-grams in the range 5-7 that occur less than 200 times.

The Linguistadors. The Linguistadors proposed a character-level CNN architecture, which was trained using ground-truth and pseudo-labels. For the first run, the model is fine-tuned using pseudo-labels for the validation set. For the second run, the model is fine-tuned using pseudo-labels for both validation and test sets. For the third run, the CNN is fine-tuned using pseudo-labels for a subset of the validation set that includes samples with 95% confidence of being correct.

Tübingen. The runs submitted by the Tübingen team (Çöltekin, 2020) consist of multiple linear SVM classifiers based on sparse character and word n-gram features, including a domain adaptation method proposed in their earlier shared task participation (Wu et al., 2019). For the first run, a base SVM model (trained only on the target development set) is first applied on the test set. Then, the model is retrained by adding the test predictions for which the classifier is confident (distance from the decision boundary is higher than 0.5) to the training set. As the final predictions, the authors take the majority vote of five classifiers trained with (slightly) different hyperparameters. For the second run, the Tübingen team used an ensemble of 20 classifiers trained on disjoint parts of the training data, while also splitting the news articles into sentences. The training data for the second submission is formed of the training and the development sets, assigning $25\times$ higher weights to tweets than to sentences taken from news articles. The third run of the Tübingen team is very similar to the second, the only difference being the filtering of the source documents based on the confidence of another classifier trained on the target development set.

UAIC. The UAIC team (Rebeja and Cristea, 2020) proposed a model based on TF-IDF encoders trained on each dialect, independently. The TF-IDF encodings are concatenated into a single tensor and provided as input to a deep learning architecture that learns to classify each data sample. The architecture is trained using categorical cross-entropy. The UAIC team submitted two runs with slightly different hyperparameters.

UPB. Similar to Anumiţi, the UPB team (Zaharia et al., 2020) submitted three runs using the Romanian BERT model (Dumitrescu et al., 2020). For the first submission, the model is trained for three epochs on text chunks of 512 tokens taken with an overlap of 128 tokens. For the second run, the Romanian BERT model is trained using an adversarial technique that alters certain examples in the data set. For the third run, the model is trained for four epochs on text chunks of 480 tokens.

SUKI. The SUKI team (Jauhiainen et al., 2020a) submitted a single run to the RDI shared task. The authors employed a custom Naïve Bayes model based on relative frequencies of character 4-grams and 5-grams as features. They removed NE tags and non-alphabetic characters from all data samples. Then, they changed the remaining characters to lowercase. The SUKI team trained the submitted model on both training and development samples.

4.3 Results

Rank	Team	Run	Method	F1 (macro)
1	Tübingen	1	SVM ensemble based on word and char n-grams	0.787592
	Tübingen	2	SVM ensemble based on word and char n-grams	0.784317
2	Anumiţi	3	SVM ensemble based on five BERT embeddings	0.775178
	Anumiţi	2	Fine-tuned uncased Romanian BERT	0.762677
	Tübingen	3	SVM ensemble based on word and char n-grams	0.756461
	Anumiţi	1	Fine-tuned cased Romanian BERT	0.746005
3	Phlyers	1	Naïve Bayes based on word n-grams	0.666090
4	SUKI	1	Naïve Bayes based on char n-grams	0.658437
	Phlyers	2	Naïve Bayes based on word n-grams	0.650884
5	UPB	1	Fine-tuned Romanian BERT	0.647577
	Phlyers	3	Naïve Bayes based on word n-grams	0.644527
	UPB	2	Fine-tuned Romanian BERT	0.563287
	UPB	3	Fine-tuned Romanian BERT	0.557496
6	UAIC	1	Deep network based on TF-IDF encodings	0.555044
	UAIC	2	Deep network based on TF-IDF encodings	0.486896
7	Akanksha	1	Character-level and phrase-level reformer	0.481325
8	The Linguistadors	2	Character-level CNN with pseudo-labels	0.429412
	The Linguistadors	3	Character-level CNN with pseudo-labels	0.411571
	The Linguistadors	1	Character-level CNN with pseudo-labels	0.396090

Table 3: Macro F_1-scores attained by the participating teams on the RDI shared task. A summary of methods and features used by participants are also included.

The runs submitted by each participant in the RDI shared task are presented in Table 3. The systems are ranked according to the macro F_1-scores. Interestingly, we observe that the top scoring system is a shallow approach based on an SVM ensemble applied on word and character n-grams. The best model, which is submitted by the Tübingen team, is closely followed by an SVM ensemble applied on fine-tuned multilingual and monolingual BERT embeddings. The results show that Tübingen and Anumiţi are the only two teams surpassing the 70% threshold. Their very good results compared with the rest of the participants are likely due to the idea of splitting the news articles in sentences. This hypothesis is also supported by the following observation. Although both Anumiţi and UPB fine-tuned the same Romanian BERT model, their results are significantly different, probably because the Anumiţi team fine-tuned the model on sentences, while UPB fine-tuned it on text chunks. Considering the domain gap between news articles and tweets, which is also caused by the high difference in the average number of tokens per sample – see (Găman and Ionescu, 2020b) – we believe that the idea of splitting the news articles into sentences to reduce the domain gap is quite useful.

4.4 Summary

In the Romanian Dialect Identification challenge, we proposed a shared task on cross-domain binary classification by dialect. A total of 8 teams participated in the competition, each submitting between 1 and 3 runs. This resulted in a total of 19 submissions, which represents an increase of almost 100% compared with last year's Moldavian vs. Romanian Cross-Dialect Topic Identification (MRC) shared

task (Zampieri et al., 2019). An interesting difference compared with the results reported for the MRC shared task is that, in the RDI shared task, the best performance is obtained by a shallow approach based on word and character n-grams. This is consistent with the results observed in previous VarDial evaluation campaigns (Zampieri et al., 2017; Zampieri et al., 2018), where some of the winners employed shallow approaches based on character n-grams (Butnaru and Ionescu, 2018; Ionescu and Butnaru, 2017). In summary, we conclude that the battle between deep and shallow approaches is still open, at least when it comes to dialect identification.

5 Social Media Variety Geolocation SMG

5.1 Datasets

The SMG task is based on three datasets from two Social Media platforms, Jodel and Twitter.

- The **BCMS subtask** is focused on geolocated tweets published in the area of Croatia, Bosnia and Herzegovina, Montenegro and Serbia in the so-called BCMS macro-language (ISO acronym HBS, code 639-3). While the independent status of the specific languages is rather disputed, there is significant variation between them.
- The **DE-AT subtask** focuses on Jodel conversations initiated in Germany and Austria, which are written in standard German but commonly contain regional and dialectal forms. Jodel is a mobile chat application that lets people anonymously talk to other users within a 10km-radius around them.
- The **CH subtask** focuses on Jodel conversations from Switzerland, which were found to be held majoritarily in Swiss German dialects. This dataset is considerably smaller, but we expect it to contain more dialect-specific cues than the DE-AT one.

The BCMS Twitter dataset is described in Ljubešić et al. (2016). The two Jodel datasets are subsets of the corpus collected by Hovy and Purschke (2018). Some additional cleaning and filtering steps have been applied to these corpora, and they have been split into training, development and test sets (see Table 4 for key figures). All three subtasks use the same data format: each instance consists of three fields, the unprocessed text of the message (BCMS) or conversation (DE-AT and CH), the latitude coordinate and the longitude coordinate. Figure 1 shows the geographic distribution of training instances.

Subtask	**Number of instances**			**Average number of tokens per instance**
	Training	**Development**	**Test**	
BCMS	320,042	39,750	39,723	13
DE-AT	336,983	46,582	48,239	71
CH	22,600	3,068	3,097	55

Table 4: SMG datasets.

5.2 Participants and Approaches

We received submissions from seven teams, with five teams participating in all three subtasks. The HeLju team submitted both unconstrained and constrained systems, whereas all other participants focused on constrained systems (i.e., not using any external training data).

The participating systems can be classified into two approaches to geolocation: a direct one which frames the problem as a double regression, and an indirect one which converts the coordinates into a finite set of dialect areas and uses a classification model to predict one of the areas.

CUBoulder-UBC. This approach is based on earlier work described in Hulden et al. (2015). It divides each geographic area into a fixed grid and uses a Naive Bayes classifier with bag-of-words features for prediction, together with kernel density estimation to avoid data sparsity. The submissions include a single system, a mean-based ensemble of 10, and a median-based ensemble of 10.

Figure 1: Geographic distribution of training instances in the three SMG datasets: DE-AT (top), CH (bottom left), BCMS (bottom right).

HeLju. The HeLju systems (Scherrer and Ljubešić, 2020) rely on the BERT architecture, where the classification output is replaced by a double regression output. For the constrained submissions (C), the BERT models are trained from scratch using the SMG training data, whereas pre-trained models are used for the unconstrained submissions (UC). The unconstrained submissions named *UC ext* include additional training data from the development set.

Piyush Mishra. This submission is based on a bidirectional LSTM that is fed with FastText embeddings (Mishra, 2020). Latitudes and longitudes are predicted by double regression with quantile loss.

SUKI. This approach divides each geographic area into a fixed grid with 81 areas and uses a n-gram language model to predict the most likely area (Jauhiainen et al., 2020a).

The Linguistadors. These submissions are based on classic regression methods (linear regression, lasso regression, and ridge regression) and rely on TF-IDF weighted input features.

UnibucKernel. The UnibucKernel team (Găman and Ionescu, 2020a) submitted two single systems, a character-level CNN (Zhang et al., 2015) with double regression output, and a Nu-SVR model trained on top of n-gram string kernels (Ionescu et al., 2016). The third system is an ensemble approach based on XGBoost, trained on the predictions provided by the two previously mentioned systems and an LSTM-based one. The LSTM is trained on top of fine-tuned German BERT embeddings.

ZHAW-InIT. The ZHAW-InIT team (Benites et al., 2020) uses unsupervised k-means clustering to infer a set of dialect classes which are then used in a classification architecture. Their systems are based

either on SVMs with TF-IDF weighted word and character n-gram features, or on the HELI language modeling architecture (ZHAW-InIT (HELI)). The SVM submission to the CH subtask (ZHAW-InIT (META)) is in fact a meta-classifier combining several SVMs with different features, whereas single SVMs are used for BCMS and DE-AT (ZHAW-InIT (SVM)).

5.3 Results

The test set predictions were evaluated on the basis of median and mean distance to the gold coordinate. Submissions are ranked by decreasing median distance, which is the official metric. For comparison, we also mention the distance values obtained from a simple centroid baseline, which predicts the center point (measured on the training data) for each test instance. Results and rankings for the three tasks are presented in Tables 5, 6, and 7 respectively. Ranks are attributed only to the best-ranked submission of each team.

Rank	Team (Run)	Median distance	Mean distance
1	HeLju (UC)	41.54	80.89
	HeLju (UC ext)	41.61	80.24
	HeLju (C)	48.99	86.83
2	ZHAW-InIT (SVM)	57.24	100.42
3	SUKI	61.01	105.11
4	CUBoulder-UBC (single)	64.76	106.67
	CUBoulder-UBC (med. ens.)	64.92	106.45
	CUBoulder-UBC (mean ens.)	66.36	102.85
5	Piyush Mishra	85.70	112.65
6	The lingustadors (Linear)	97.16	141.88
	The lingustadors (Ridge)	105.54	141.58
	The lingustadors (Lasso)	107.04	145.68
	Centroid baseline	*107.10*	*145.72*
	ZHAW-InIT (HELI)	111.40	130.23

Table 5: SMG shared task - BCMS results. Unconstrained submissions above the horizontal line, constrained ones below.

Rank	Team (Run)	Median distance	Mean distance	Dialect area accuracy
1	HeLju (UC ext)	143.30	166.64	36.1%
	HeLju (UC)	143.85	168.45	34.8%
	HeLju (C)	159.59	183.97	29.5%
2	Piyush Mishra	183.99	204.93	21.8%
3	CUBoulder-UBC (mean ens.)	198.27	218.51	25.1%
	Centroid baseline	*201.34*	*221.55*	*17.7%*
4	ZHAW-InIT (SVM)	205.81	230.78	27.7%
	CUBoulder-UBC (median ens.)	214.72	235.62	25.4%
	ZHAW-InIT (HELI)	217.80	241.33	19.5%
	CUBoulder-UBC (single)	219.08	239.47	25.7%
5	SUKI	243.12	266.85	24.4%

Table 6: SMG shared task - DE-AT results. Unconstrained submissions above the horizontal line, constrained ones below.

For the DE-AT and CH subtasks, we also provide a dialect area accuracy measure. These are based on partitions of the areas into different dialectal areas based on previous dialectological research (Lameli,

Rank	Team (Run)	Median distance	Mean distance	Dialect area accuracy
1	HeLju (UC ext)	15.45	22.45	72.6%
	HeLju (UC)	15.72	22.67	72.9%
2	ZHAW-InIT (META)	15.93	25.06	72.6%
	ZHAW-InIT (HELI)	17.66	26.21	69.4%
	HeLju (C)	17.97	26.04	68.7%
3	CUBoulder-UBC (mean ens.)	19.49	27.63	66.9%
	CUBoulder-UBC (median ens.)	19.66	28.83	66.4%
	CUBoulder-UBC (single)	19.99	29.09	66.2%
4	SUKI	23.96	34.59	57.0%
5	UnibucKernel (ens.)	25.57	30.52	53.9%
6	The Linguistadors (Ridge)	26.70	31.21	50.3%
	UnibucKernel (SVR)	26.78	31.49	51.1%
7	Piyush Mishra	27.31	33.20	53.3%
	The Linguistadors (Linear)	35.70	39.66	33.6%
	UnibucKernel (CNN)	40.23	42.87	29.8%
	Centroid baseline	*41.38*	*48.16*	*15.8%*
	The Linguistadors (Lasso)	41.60	48.19	15.8%

Table 7: CH results. Unconstrained submissions above the horizontal line, constrained ones below.

2013; Scherrer and Stoeckle, 2016).[5] Dialect area accuracy represents the percentage of test instances whose predicted coordinates lie inside the same area as the gold coordinates.

Rather unsurprisingly, the unconstrained approaches outperform the constrained ones, but only by a small margin in the CH subtask. There is no clear winning approach among the constrained submissions. BERT (HeLju) works well in large-data settings, but underperforms on the CH subtask where classical approaches are more competitive. The ZHAW-InIT and CUBoulder-UBC systems show that a classification strategy with a fixed set of classes can outperform a regression strategy that learns to predict longitudes and latitudes directly. This finding may be due to the fact that social media posts are not randomly scattered across space, but tend to gather around a relatively small number of larger cities and agglomerations.

More generally, the DE-AT subtask has turned out to be the hardest one: only half of the submitted systems managed to beat the baseline, and unlike in the other subtasks, no system managed to halve the baseline distance. This suggests that the regional features in the DE-AT Jodel corpus are too sparse to be learned reliably.

In terms of evaluation measures, the median and mean distances are highly correlated. Dialect area accuracy also yields a similar picture overall, but some differences are noteworthy. For DE-AT, all systems clearly outperform the baseline on this measure, and the ZHAW-InIT (SVM) submission turns out to be much more competitive than the distance measures suggest. This submission confirms its advantage on the CH task, where it is indistinguishable from the unconstrained submissions in terms of dialect area accuracy.

5.4 Summary

For the first time at VarDial, we have proposed a dialect geolocation task, in which the prediction outputs are coordinate pairs rather than variety labels. The SMG task attracted a total of seven participants across three subtasks, a number that is comparable with other VarDial tasks in recent years. We received a wide range of technical solutions: solutions based on deep learning as well as traditional machine learning, constrained as well as unconstrained solutions, and regression-based as well as classification-

[5]The DE-AT areas are based on those used in (Hovy and Purschke, 2018), augmented with a single area covering the territory of Austria. The CH areas correspond to the 10-cluster solution presented in (Scherrer and Stoeckle, 2016).

based approaches. The best scores were obtained by unconstrained solutions based on pre-trained BERT models, on all three subtasks. Thanks to its reliance on easily available geolocated messages from social media services, another edition of the SMG task could be envisaged, possibly focusing on different language areas.

6 Uralic Language Identification ULI

The first edition of the ULI shared task was a language identification task focusing on differentiating between minority Uralic languages and distinguishing them from a large number of other languages.

We define minority Uralic languages as those languages that are not official state languages in the countries where they are spoken. This definition excludes Estonian, Finnish and Hungarian. The remaining Uralic languages are all endangered. They can also be characterized as extremely diverse, at least when it comes to their use and current situation. Some of the languages in the shared task are extinct, while others have very young and varying orthographies that are still becoming established. Nevertheless, the task also includes languages that are widely used in the modern society and have a large online presence. Most of the Uralic languages spoken in Russia are written using the Cyrillic alphabet, often with additional individual characters that differ from the character set used for Russian. Since the Uralic languages form a large and old language family, the varieties in the task are generally far apart from one another. At the same time, the task also contains closely related Uralic languages from individual branches, which share a large percentage of their vocabulary and features.

The shared task included a total of 178 languages, of which 29 were Uralic minority languages. The 29 endangered Uralic languages were considered relevant and the 148 languages non-relevant. The ULI task consisted of three tracks using the same training and testing data. The tracks differed from each other in how they were scored.

The motivation behind including the non-relevant languages in the shared task was to simulate the situation we faced when we were automatically searching for minority Uralic languages on the Internet during the Finno-Ugric Languages and the Internet project (Jauhiainen et al., 2015). The different ways of scoring the tracks was also designed to highlight the inherent difficulties of such a search. The third track did not especially focus on the relevant languages, the second track focused on the relevant languages as a group, and the first track forced the participants to consider even the most rare of the relevant languages.

The first track, ULI-RLE (Relevant languages as equals), considered all the relevant languages equal in value and the aim was to maximize their average F-score. This is important when one is interested in finding rare languages on, for example, the Internet. The F-score was calculated as a macro-averaged F1 score over the relevant languages in the training set.

The second track, ULI-RSS (Relevant sentences as equals), considered each sentence in the test set that was written in or was predicted to be in a relevant language as equals. When compared with the first track, this track gave less importance to the very rare languages as their precision was not as important when the resulting F-score was calculated. The resulting F-score was calculated as a micro-F1 over the sentences in the test set for both the sentences in the relevant languages and the ones that were predicted to be in relevant languages.

In the first two tracks, there was no difference between the non-relevant languages. All the non-relevant languages could have been labeled as English in the submissions and it would not have changed the resulting F1-scores. The third track, ULI-178 (All 178 languages as equals), however, did not focus on the 29 relevant languages, but instead the target was to maximize the average F-score over all the 178 languages present in the training set. The ULI shared task, and especially this track, was the language identification shared task with the largest number of languages used so far. The F-score was calculated as a macro-F1 score over all the languages in the training set.

6.1 Dataset

For training, we provided texts from the Wanca 2016 corpora (Jauhiainen et al., 2019a) for the relevant languages and from the Leipzig corpora collection (Goldhahn et al., 2012) for the non-relevant languages.

The number of lines for the non-relevant languages in the training data varied from 10,000 lines of Cebuano and Corsican to 3 million lines of Indonesian. As relevant language sentences in the test set from the forthcoming Wanca 2017 corpora (Jauhiainen et al., 2020b), we chose those sentences that were not present in the Wanca 2016 corpora which have been published in the Language Bank of Finland. The sentences for the non-relevant languages were from the Leipzig corpora collection. We did not create a separate set for development so the participants had to decide themselves how to use the given training material for that as well. The dataset used in the ULI shared task, as well as its creation, is described in detail by Jauhiainen et al. (2020b).

6.2 Participants and Approaches

Unfortunately, the ULI shared task had only one team submitting results to the tracks. The NRC team submitted three runs for each of the shared task tracks. All the runs used BERT-related deep neural networks taking sequences of characters as input similar to what the NRC team used when they won the CLI shared task (Jauhiainen et al., 2019b) in the previous VarDial Evaluation Campaign (Bernier-Colborne et al., 2019). The encoders of the networks were pre-trained on masked language modeling (MLM) and sentence pair classification (SPC) tasks (Devlin et al., 2019). The third run on each track was using only the information on the training set as opposed to the second run, in which the MLM was also done on the unlabeled test set in order to adapt the model. The first run on each track was a plurality voting ensemble of the six models used in the second and third runs of all the tracks.

6.3 Results

For the baseline, we used an implementation of the HeLI method equal to the one we used when evaluating language identification methods for 285 languages (Jauhiainen et al., 2017). The baseline and the NRC teams results are listed in Tables 8, 9, and 10.

Rank	**Team**	**Run**	**Method**	**Relevant macro F_1**
	baseline		HeLI	0.8004
1	NRC	2	deep neural network with adaptation to the test set	0.2996
	NRC	1	ensemble of 6 deep neural networks	0.2872
	NRC	3	deep neural network	0.2514

Table 8: ULI shaed task - RLE results.

Rank	**Team**	**Run**	**Method**	**Relevant micro F_1**
	baseline		HeLI	0.9632
1	NRC	1	ensemble of 6 deep neural networks	0.2596
	NRC	2	deep neural network with adaptation to the test set	0.1547
	NRC	3	deep neural network	0.1359

Table 9: ULI shared task - RSS results.

Rank	**Team**	**Run**	**Method**	**Macro F_1**
	baseline		HeLI	0.9252
1	NRC	2	deep neural network with adaptation to the test set	0.6751
	NRC	3	deep neural network	0.6628
	NRC	1	ensemble of 6 deep neural networks	0.6356

Table 10: ULIshared task - 178 results.

All the results submitted by the NRC team are well below the baselines. After the shared task results

were announced, the NRC team investigated reasons for the low performance of their classifiers and found that the low scores were mostly due to a flaw in the function they used to sample the data for training and evaluation (Bernier-Colborne and Goutte, 2020).

6.4 Summary

Needless to say, we were not happy that the ULI task attracted the submissions of only one team. The shared tasks at VarDial have historically attracted a good number of submissions but, as previously mentioned, the VarDial Evaluation Campaign 2020 run during the early stages of the COVID-19 pandemic, a period in which significant disruption has been observed in universities and research centers worldwide. This is likely to have precluded more teams from participating. Furthermore, we acknowledge that we did not make the task easy to participate with the larger than normal training sets. The results of the participating team also suggest that the task might have been more difficult than we anticipated. Due to the low number of participants in the shared task and the challenges caused by the COVID-19 pandemic, we have decided to continue accepting submissions until the next edition of the ULI shared task. Thus, we are not yet publishing the gold-labeled test set. Instead, we will set up a web-page[6] with information on how to request the training and the test sets. The web-page will also feature a table with all the results submitted so far.

7 Conclusion

In this paper we present the results and findings of the shared tasks organized as part of the VarDial Evaluation Campaign 2020. Three shared tasks were organized this year: Romanian Dialect Identification (RDI), Social Media Variety Geolocation (SMG), and Uralic Language Identification (ULI). Each of these tasks tackled an important challenge in language and dialect identification on different languages and dialects. Furthermore, in these tasks we provided participants with new datasets that will be made freely available to the research community after the competitions.

A total of 14 teams submitted runs across the three shared tasks. We included short descriptions for each team's systems in this report and references to all 11 system description papers in Table 1. A complete description of these systems is available in the system description papers published in the VarDial workshop proceedings.

Acknowledgments

We would like to thank the shared task participants for their participation, support, and the feedback provided. We further thank the VarDial program committee for reviewing all submissions.

References

Fernando Benites, Manuela Hürlimann, Pius von Däniken, and Mark Cieliebak. 2020. ZHAW-InIT - Social Media Geolocation at VarDial 2020. In *Proceedings of VarDial*.

Gabriel Bernier-Colborne and Cyril Goutte. 2020. Challenges in neural language identification: NRC at VarDial 2020. In *Proceedings of VarDial*.

Gabriel Bernier-Colborne, Cyril Goutte, and Serge Léger. 2019. Improving cuneiform language identification with BERT. In *Proceedings of VarDial*.

Andrei M. Butnaru and Radu Ionescu. 2018. UnibucKernel Reloaded: First Place in Arabic Dialect Identification for the Second Year in a Row. In *Proceedings of VarDial*.

Andrei M. Butnaru and Radu Tudor Ionescu. 2019. MOROCO: The Moldavian and Romanian Dialectal Corpus. In *Proceedings of ACL*.

Andrea Ceolin and Hong Zhang. 2020. Discriminating between standard Romanian and Moldavian tweets using filtered character ngrams. In *Proceedings of VarDial*.

[6] http://urn.fi/urn:nbn:fi:lb-2020102201

Çağrı Çöltekin. 2020. Dialect identification under domain shift: Experiments with discriminating Romanian and Moldavian. In *Proceedings of VarDial*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Ştefan Daniel Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of EMNLP*.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of LREC*.

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

Mihaela Găman and Radu Tudor Ionescu. 2020a. Combining deep learning and string kernels for the localization of Swiss German tweets. In *Proceedings of VarDial*.

Mihaela Găman and Radu Tudor Ionescu. 2020b. The Unreasonable Effectiveness of Machine Learning in Moldavian versus Romanian Dialect Identification. *arXiv preprint arXiv:2007.15700*.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of EMNLP*.

Mans Hulden, Miikka Silfverberg, and Jerid Francom. 2015. Kernel density estimation for text-based geolocation. In *Proceedings of AAAI*.

Radu Tudor Ionescu and Andrei Butnaru. 2017. Learning to identify Arabic and German dialects using multiple kernels. In *Proceedings of VarDial*.

Radu Tudor Ionescu, Marius Popescu, and Aoife Cahill. 2016. String kernels for native language identification: Insights from behind the curtains. *Computational Linguistics*, 42(3):491–525.

Heidi Jauhiainen, Tommi Jauhiainen, and Krister Lindén. 2015. The Finno-Ugric Languages and The Internet Project. In *Proceedings of IWCLUL*.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2017. Evaluation of Language Identification Methods Using 285 Languages. In *Proceedings of NoDaLiDa*.

Heidi Jauhiainen, Tommi Jauhiainen, and Krister Linden. 2019a. Wanca in Korp: Text corpora for underresourced Uralic languages. In *Proceedings of RDHUM*.

Tommi Jauhiainen, Heidi Jauhiainen, Tero Alstola, and Krister Lindén. 2019b. Language and dialect identification of cuneiform texts. In *Proceedings of VarDial*.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019c. Automatic Language Identification in Texts: A Survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2020a. Experiments in language variety geolocation and dialect identification. In *Proceedings of VarDial*.

Tommi Jauhiainen, Heidi Jauhiainen, Niko Partanen, and Krister Lindén. 2020b. Uralic Language Identification (ULI) 2020 shared task dataset and the Wanca 2017 corpora. In *Proceedings of VarDial*.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2019. Reformer: The Efficient Transformer. In *Proceedings of ICLR*.

Ritesh Kumar, Bornini Lahiri, Deepak Alok, Atul Kr. Ojha, Mayank Jain, Abdul Basit, and Yogesh Dawar. 2018. Automatic Identification of Closely-related Indian Languages: Resources and Experiments. In *Proceedings of LREC*.

Alfred Lameli. 2013. *Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*. Walter de Gruyter.

Nikola Ljubešić, Tanja Samardžić, and Curdin Derungs. 2016. TweetGeo - a tool for collecting, processing and analysing geo-encoded linguistic data. In *Proceedings of COLING*.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, 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 VarDial*.

Piyush Mishra. 2020. Geolocation of tweets with a BiLSTM regression model. In *Proceedings of VarDial*.

Cristian Popa and Vlad Ștefănescu. 2020. Applying multilingual and monolingual Transformer-based models for dialect identification. In *Proceedings of VarDial*.

Petru Rebeja and Dan Cristea. 2020. A dual-encoding system for dialect classification. In *Proceedings of VarDial*.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. ArchiMob–A corpus of spoken Swiss German. In *Proceedings of LREC*.

Yves Scherrer and Nikola Ljubešić. 2020. HeLju@VarDial 2020: Social media variety geolocation with BERT models. In *Proceedings of VarDial*.

Yves Scherrer and Philipp Stoeckle. 2016. A quantitative approach to Swiss German – Dialectometric analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica*, 1(24):92–125.

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 BUCC*.

Nianheng Wu, Eric DeMattos, Kwok Him So, Pin-zhen Chen, and Çağrı Çöltekin. 2019. Language discrimination and transfer learning for similar languages: Experiments with feature combinations and adaptation. In *Proceedings of VarDial*.

George-Eduard Zaharia, Andrei-Marius Avram, Dumitru-Clementin Cercel, and Traian Rebedea. 2020. Exploring the power of Romanian BERT for dialect identification. In *Proceedings of VarDial*.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial Evaluation Campaign 2017. In *Proceedings of VarDial*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Dirk Speelman, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. Language Identification and Morphosyntactic Tagging: The Second VarDial Evaluation Campaign. In *Proceedings of VarDial*.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei Butnaru, and Tommi Jauhiainen. 2019. A Report on the Third VarDial Evaluation Campaign. In *Proceedings of VarDial*.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural Language Processing for Similar Languages, Varieties, and Dialects: A Survey. *Natural Language Engineering*, 26:595–612.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Proceedings of NIPS*, pages 649–657.