Findings of the VarDial Evaluation Campaign 2022

Aepli, Noëmi

COLING
2022-10

Aepli , N , Anastasopoulos , A , Chifu , A-G , Domingues , W , Faisal , F , Gaman , M , Ionescu , R T & Scherrer , Y 2022 , Findings of the VarDial Evaluation Campaign 2022 . in Y þÿ S c h e r r e r , T J a u h i a i n e n , N L j u b e a i , P N a k o v , J T i e d e m a n n & M Z a m Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects : The 29th International Conference on Computational Linguistics . International conference on computational linguistics , no. 19 , vol. 29 , COLING , Gyeongju , pp. 1-13 , Workshop on NLP for Similar Languages, Varieties and Dialects , Gyeongju , Korea, Republic of , 16/10/2022 .

http://hdl.handle.net/10138/350298

Findings of the VarDial Evaluation Campaign 2022

Noëmi Aepli[1] · ITDI, **Antonios Anastasopoulos**[2] · DialQA, **Adrian Chifu**[3] · FDI, **William Domingues**[3] · FDI, **Fahim Faisal**[2] · DialQA, **Mihaela Găman**[4] · FDI, **Radu Tudor Ionescu**[4] · FDI, **Yves Scherrer**[5] · ITDI

[1]University of Zurich, [2]George Mason University, [3]Aix-Marseille Université, [4]University of Bucharest, [5]University of Helsinki

Abstract

This report presents the results of the shared tasks organized as part of the VarDial Evaluation Campaign 2022. The campaign is part of the ninth workshop on Natural Language Processing (NLP) for Similar Languages, Varieties and Dialects (VarDial), co-located with COLING 2022. Three separate shared tasks were included this year: Identification of Languages and Dialects of Italy (ITDI), French Cross-Domain Dialect Identification (FDI), and Dialectal Extractive Question Answering (DialQA). All three tasks were organized for the first time this year.

1 Introduction

The workshop series on *NLP for Similar Languages, Varieties and Dialects* (VarDial), traditionally co-located with international conferences, has reached its ninth edition. Since the first edition, VarDial has hosted shared tasks on various topics such as language and dialect identification, morphosyntactic tagging, question answering, and cross-lingual dependency parsing. The shared tasks have featured many languages and dialects from different families and data from various sources, genres, and domains (Chakravarthi et al., 2021; Gaman et al., 2020; Zampieri et al., 2019, 2018, 2017; Malmasi et al., 2016; Zampieri et al., 2015, 2014).

We offered three shared tasks as part of the VarDial Evaluation Campaign 2022, which we present in this paper: Identification of Languages and Dialects of Italy (ITDI), French Cross-Domain Dialect Identification (FDI), and Dialectal Extractive Question Answering (DialQA).

This overview paper is structured as follows: in Section 2, we briefly introduce the three shared tasks. Section 3 presents the teams that submitted systems to the shared tasks. Each task is then discussed in detail, focusing on the data, the participants' approaches, and the obtained results. Section 4 is dedicated to ITDI, Section 5 to FDI, and Section 6 to DialQA.

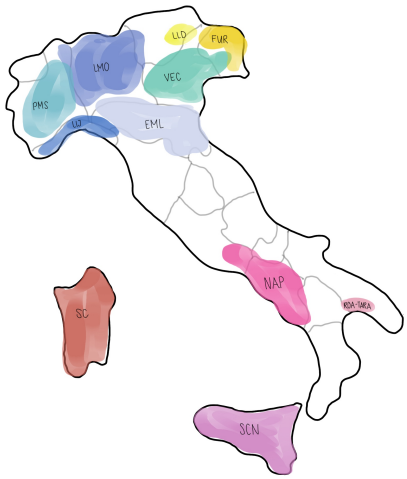


Figure 1: Rough regions where the eleven considered languages and dialects of Italy are spoken. magenta: Italo-Dalmatian; turquoise: Gallo-Italian; yellow: Gallo-Rhaetian; red: Sardinian. The map is vague; the situation is more complex. However, it gives an idea of where in Italy to locate the varieties.

2 Shared Tasks at VarDial 2022

2.1 Identification of Languages and Dialects of Italy (ITDI)

Italy features a rich linguistic diversity with numerous local and regional language varieties. Many of the varieties form a continuum, but some others are very distinct. The ITDI shared task focuses on eleven language varieties that belong to the Romance language branch (like Italy's official language, Italian) and have their own Wikipedia.[1] Figure 2 displays the relations of the eleven language varieties according to the classification by Ethnologue (Eberhard et al., 2022), and Figure 1 shows the approximate regions where they are mainly spo-

[1]by March 2022, when we created the shared task.

Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 1–13

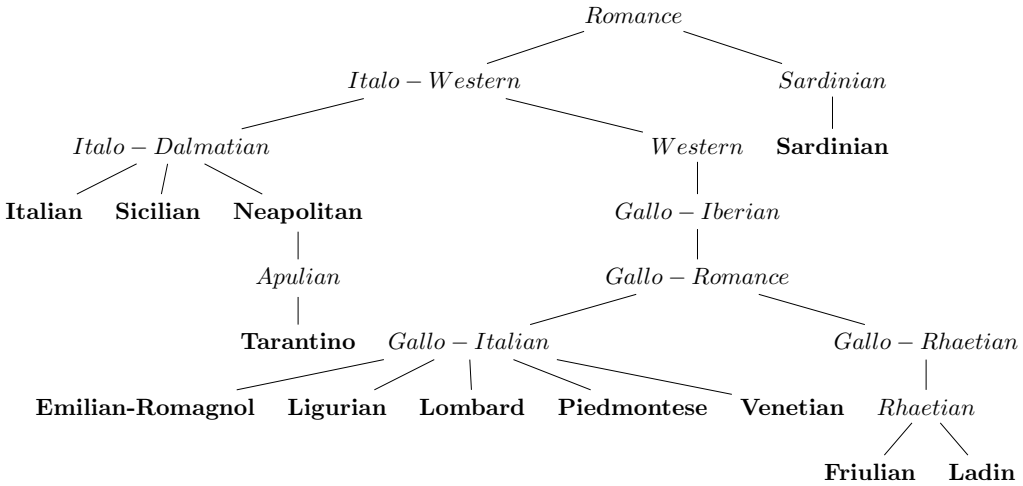


Figure 2: Relations between the eleven considered languages and dialects of Italy, according to Ethnologue.

ken.[2] More fine-grained classifications within dialects are possible. We must remember that classification into categories is imprecise for a continuum as we work with distinct rather than continuous values. Depending on the availability of data, all the data splits (training, development, test) may contain one or several sub-varieties of the category predetermined by the Wikipedia dumps. Furthermore, we rely on the categorization by the authors of the texts, which might not be the one every native speaker agrees upon.

To the best of our knowledge, no previous language identification research focuses exclusively on Italy's languages and dialects. However, some of the language varieties featured in our shared task have been part of other research related to language identification. Jauhiainen et al. (2022) present a detailed overview. More generally, Ramponi (2022) reviews recent work on NLP for the language varieties of Italy and identifies the most pressing challenges for their computational processing.

The ITDI task is a cross-domain classification task in which the model is required to discriminate between eleven languages and dialects of Italy. The setting is similar to a real-world problem because the training data consists only of Wikipedia dumps, i.e., careful pre-processing is part of the task. Furthermore, the data is not balanced in any of the data splits. Finally, development and test splits only contain sentences of distinct subsets of the eleven languages and dialects and come from different sources and domains (see Appendix A.1 for details). The submission format is closed, meaning that participants cannot use additional data to train their models – exceptions are off-the-shelf pre-trained language models, which only one team made use of.

[2]We created this map according to https://upload.wikimedia.org/wikipedia/commons/3/32/Dialetti_e_lingue_in_Italia.png (Antonio Ciccolella via Wikimedia Commons, 2015).

2.2 French Cross-Domain Dialect Identification (FDI)

For the 2022 French Cross-Domain Dialect Identification (FDI) shared task, participants had to train a model on news samples collected from a set of publication sources and evaluate it on news samples collected from a different set of publication sources. To ensure that dialect identification models do not rely on features such as author style or text topic, the publication sources and the topics are different across splits. Therefore, participants had to build a model for a cross-domain four-way classification by dialect task, in which a classification method is required to discriminate between the French (FR), Swiss (CH), Belgian (BE), and Canadian (CA) dialects observed in news samples. For the shared task, we provided participants with the French Cross-Domain Dialect dataset (Găman et al., 2022), which contains French, Swiss, Belgian, and Canadian samples of text collected from the news domain. The corpus is divided into training, validation and test, such that the training set contains 358,787 samples, the development set 18,002 samples, and the test set 36,733 samples.

Participants are evaluated in two separate scenarios: open and closed. In the closed format, participants are not allowed to use pre-trained models or external data to train their models. In the open

Team	ITDI	FDI	DialQA	System Description Paper
DCT		✓		Gillin (2022)
ETHZ	✓			Camposampiero et al. (2022)
NRC		✓		Bernier-Colborne et al. (2022)
Phlyers	✓			Ceolin (2022)
SUKI	✓	✓		Jauhiainen et al. (2022)

Table 1: The teams that participated in the VarDial Evaluation Campaign 2022.

format, participants are allowed to use external resources such as unlabeled corpora, lexicons, and pre-trained embeddings (e.g. CamemBERT (Martin et al., 2020)), but the use of additional labeled data is still not allowed.

2.3 Dialectal Extractive Question Answering (DialQA)

Question Answering (QA) systems are capable of answering human prompts with or without context. With the advancement of query-based smartphone assistants (eg. Google Assistant, Amazon Alexa, or Apple Siri), the use-case scenarios of such systems have already reached a global scale. However, in most cases, the traditional text-based extractive QA systems still follow the training routine on error-free written text, whereas the real-world scenario contains error-prone interfaces.

This year we introduced the DialQA shared task to build QA systems that are robust to dialectal variation. To make extractive QA systems more representative of real-world scenarios, we prepared an evaluation dataset based on the existing TyDi-QA (Clark et al., 2020) dataset with two additional dimensions. First, the augmented question text contains dialectal and/or geographical language variations. Second, we provide these questions in spoken form to match the scenario of users querying virtual assistants for information. The participants could either (a) use the baseline automatic speech recognition outputs for each dialect to make a robust text-based QA system, or (b) they may use the provided audio recordings of the questions to make a dialect-robust ASR system which can be then evaluated with a baseline QA system, or (c) both of the above. The shared task was based on the SD-QA (Faisal et al., 2021) development and test datasets for English, Arabic, and Kiswahili varieties, as well as code for training text-based baseline extractive QA systems based on TyDi-QA.

3 Participating Teams

A total of five teams submitted runs to the ITDI and FDI shared tasks. Unfortunately, we did not receive any submissions for DialQA. In Table 1, we list the teams that participated in the shared tasks, including references to the system description papers which are published as parts of the VarDial workshop proceedings. Detailed information about the submissions is included in the task-specific sections below.

4 Identification of Languages and Dialects of Italy

4.1 Dataset

The training set consists of eleven Wikipedia dumps:[3] Emilian-Romagnol (EML), Friulian (FUR), Ladin (LLD), Ligurian (LIJ), Lombard (LMO), Neapolitan (NAP), Piedmontese (PMS), Sardinian (SC), Sicilian (SCN), Tarantino (ROA_TARA) and Venetian (VEC). We provided the participants with a script to download and extract the dumps on the basis of WikiExtractor (Attardi, 2015).

The development and test sets come from several online sources.[4] We only included sentences with a minimum length of five and a maximum of 35 tokens. Table 2 shows the number of articles (training set) and sentences (development and test set) of the data splits. The released test set contains 11,090 lines.[5]

4.2 Participants and Approaches

ETHZ: The predictions submitted by the ETHZ team (Camposampiero et al., 2022) were produced by a logistic regression (using a sag solver and

[3] `pages-articles-multistream.xml.bz2`, from 01.03.2022, now available on GitHub: https://github.com/noe-eva/ITDI_2022.

[4] See Appendix A.1 and for more information.

[5] Including three empty lines, which we deleted for the evaluation.

Language	Tag	Train articles	Dev sentences	Test sentences
Emilian-Romagnol	EML	12,996	–	825
Friulian	FUR	3,750	676	1,323
Ladin	LLD	11,981	–	2,200
Ligurian	LIJ	10,912	617	2,282
Lombard	LMO	50,518	1,231	689
Neapolitan	NAP	14,789	–	2,026
Piedmontese	PMS	66,268	1,191	–
Sardinian	SC	7,419	477	–
Sicilian	SCN	26,464	1,371	–
Tarantino	ROA_TARA	9,322	–	603
Venetian	VEC	68,955	1,236	1,139
Total		283,374	6,799	11,087

Table 2: Number of articles (train) and sentences (dev/test) in the ITDI data set.

class weights) and a BERT model built on the `dbmdz-xxl-cased`[6] model. The logistic regression model ended up in fifth place. The team improved the model by a better choice of class weights but it was not considered in the ranking because it was a late submission. The BERT model brought up the rear of the team submissions.

Phlyers: The Phlyers (Ceolin, 2022) submitted three runs based on deep feedforward neural networks (DNN). The team mainly used the development data for training where possible and Wikipedia data only for the language varieties not present in the development set. For the first submission, the team re-trained the DNN, excluding PMS and ROA_TARA. The second and third submissions were similar but re-trained using the label/sentences from the test set for which the predicted label was associated with a high likelihood (with different thresholds for the two submissions), following a language model adaptation strategy.

SUKI: The SUKI team (Jauhiainen et al., 2022) applied the system they used for the FDI shared task (see Section 5.2), which is also the system they used in their winning submission of the 2021 edition of Romanian Dialect Identification (Jauhiainen et al., 2021). It is a Naïve Bayes-based method using the observed relative frequencies of multiple size character n-grams as probabilities. The system uses an adaptation technique to learn from the test data. The three submissions mainly differ in the training data used. The first submission used combined training and development data, and the second just the training data. The third system combined the training and development data, leaving out the data for PMS and SC because the number of instances did not meet their threshold.

For the ITDI shared task, the SUKI team used their own method to extract the training data from the dumps and performed extensive filtering and pre-processing, making use of their extensive experience with Wikipedia data.

Baselines: We created three baselines. The weakest one (Baseline 1) with a weighted F1-score of 0.1322 shows the results of applying an off-the-shelf tool for language identification: `FastText`[7] (Joulin et al., 2016b,a). Note that this model has been trained on earlier Wikipedia dumps and only supports seven of our eleven languages but not Friulian, Ladin, Ligurian, and Tarantino. We created this baseline by considering the ten best predictions for each sentence and took the first prediction that was one of the eight remaining varieties.

For the two other baselines (Baseline 2 and Baseline 3), we trained Support Vector Machines (SVM) with TF-IDF features using the `scikit-learn` toolkit (Pedregosa et al., 2011). We used the training data as is, i.e., no pre-processing was done after extracting the dumps except splitting the text at the line breaks (`\n`) produced by the extraction script. Baseline 2 was trained with character unigrams. It was mainly intended to see whether some individual characters are specific to certain

[6] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

[7] https://dl.fbaipublicfiles.com/fasttext/supervised-models/lid.176.bin

Team rank	Submission rank	Team	Run	Weighted-F1	Macro-F1
1	1	**SUKI**	2	0.9007	0.6729
	2	SUKI	1	0.8983	0.6714
	3	SUKI	3	0.8982	0.7458
	–	Organizers	Baseline 3	0.7726	0.5193
	*	ETHZ		0.7058	0.4885
2	4	**Phlyers**	3	0.6943	0.5379
3	5	**ETHZ**	2	0.6880	0.4828
	6	Phlyers	1	0.6631	0.5188
	7	Phlyers	2	0.6365	0.5094
	8	ETHZ	1	0.5760	0.4224
	–	Organizers	Baseline 2	0.4899	0.3424
	–	Organizers	Baseline 1	0.1322	0.1004

Table 3: Ranking of the teams and submissions according to the weighted average F1-score. The * marks a late submission by team ETHZ, which is not ranked. The baselines were created by the shared task organizers.

language varieties. It resulted in a weighted F1-score of 0.4899 and was beaten by all the submissions. The second SVM was trained on character 1-to-4-grams. It reached a weighted F1-score of 0.7726 and was only outperformed by the three submissions of team SUKI.

4.3 Results

The submissions were ranked according to the weighted average F1-score. Table 3 presents the ranking of the submissions and baselines. For comparison, we also report the macro-averaged F1-scores. We got one late submission which is marked by * in the table.

With three top-ranked submissions, team SUKI is a clear winner with their Naïve Bayes-based method using an adaptation technique. The team in the second place is Phlyers with one of their DNN models, closely followed by the logistic regression system by the ETHZ team in the third place.

One salient result was a very low recall for Ligurian by team Phlyers, which came with the cost of a few percentage points in the F1-score because Ligurian was heavily weighted with many sentences in the test set. This underprediction is due to their strategy to use mainly the development set for the varieties included in the development data, which did not work out well for this setting because apparently, the Wikipedia training data was closer to the one Wikisource book of the test set than the other Wikisource books in the development set.

The gap between the first team and the other submissions is quite big. The reason seems to be the sum of different optimal choices, like a more extensive pre-processing and the use of adaptive language models. The fourth-ranked system by Phlyers also used adaptive language models but had a different data strategy, while the third baseline ranking in between them was created without any pre-processing of the data.

Figure 3 displays the confusion matrices of the three baselines. Tarantino is the dialect with which all the systems struggled. In the best baseline (Figure 3c) and all the team submissions, it mostly gets confused with Neapolitan and Sicilian, which makes sense considering the relations in Figure 2 where Tarantino is a sub-dialect of Neapolitan further down in the language tree. Furthermore, looking at the best baseline, Neapolitan was often classified as Sicilian; Emilian-Romagnol and Venetian as Lombard; and Ladin as Venetian. The first two pairs are in the same subgroup (Gallo-Italian), while the latter pair is not so closely related in the language tree but geographically close, which might explain overlapping features of some kind in the used data set. In addition, most of the development and test data comes from Wikisource and websites, both of which have specific features; older texts for the former and texts most likely written by several and younger users for the latter. The Friulian data comes from a book (dev) and a newspaper (test) which can be considered as "controlled" in the aforementioned aspects.

Looking at Figure 3a, we have to keep in mind that FastText does not include Friulian, Ladin, Ligurian, or Tarantino. Lombard, Neapolitan, and Emilian-Romagnol seem the easiest to classify,

while Friulian gets mostly misclassified with one other variety that is linguistically unrelated. The other varieties have very high entropy and were often classified as `unk`, i.e., something other than the eight included language varieties.

4.4 Summary

We proposed a closed cross-domain classification task for the Identification of Languages and Dialects of Italy shared task. We received a total of nine submissions[8] coming from three different teams. The results of the submissions are distributed over a wide range from 0.5760 to 0.9007 weighted F1-score, with two baselines even worse.

Furthermore, the differences between the results of the eleven language varieties are enormous, probably for several reasons. As data used in this shared task comes from many different sources, there are several factors to consider: different genres, domains, writing styles, average sentence length, number of authors (each with their own style), and year of publication, to name but a few.

Unsurprisingly, an off-the-shelf system like FastText performs quite poorly for language varieties, even those included in its training data. However, a shallow machine learning system like Naïve Bayes, support vector machine, or logistic regression can achieve good performance for most language varieties included in this task.

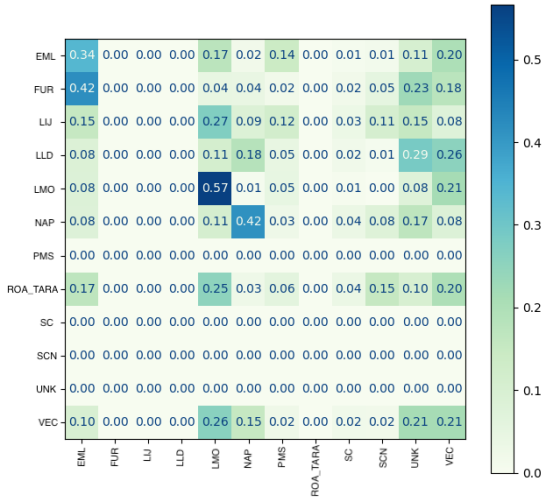
Along with this shared task, we release a newly collected and annotated data set for language identification featuring the previously mentioned eleven languages and dialects of Italy. The shared task and data are available on GitHub: https://github.com/noe-eva/ITDI_2022.

5 French Cross-Domain Dialect Identification

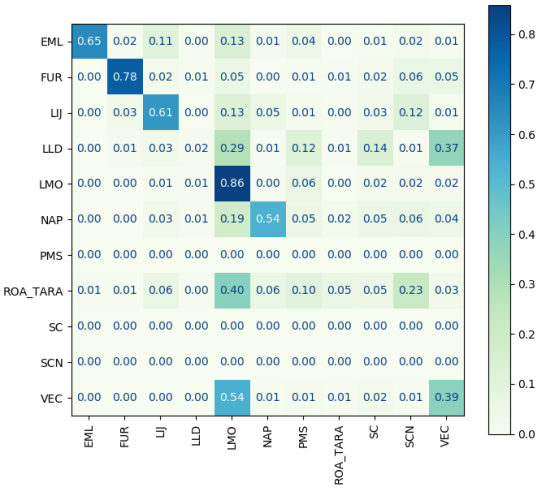
5.1 Dataset

The French Cross-Domain (FreCDo) corpus (Găman et al., 2022) contains plain text excerpts from news samples collected from public news websites in France, Switzerland, Belgium, and Canada. The corpus is divided into training, validation, and test, such that the publication sources and topics are distinct across splits. The corpus evaluates the models' ability to solve a cross-domain four-way dialect classification task. The text samples are pre-processed to hide named entities, thus eliminating country-specific clues. The named entities

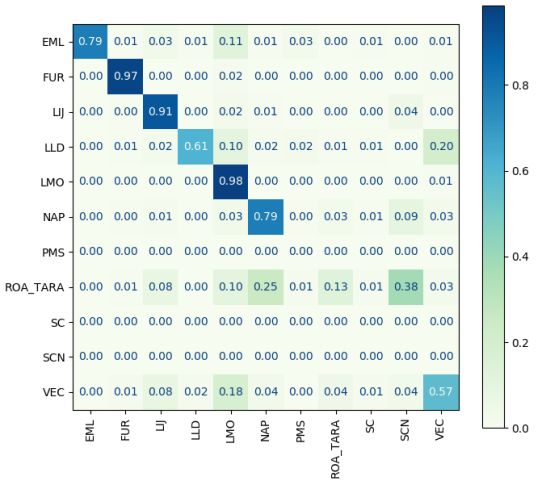
[8] one of which was a late submission



(a) Baseline 1: FastText



(b) Baseline 2: Unigram SVM



(c) N-Gram SVM

Figure 3: Confusion matrices of the three baselines (see Section 4.2). The numbers indicate the counts normalized over the true conditions of the test set (i.e. no instances of PMS, SC, and SCN in the gold standard). True labels on the y-axis, predicted on the x-axis.

Split	Country	# Samples	# Tokens
Train	BE	121,746	11,619,874
	CA	34,003	2,505,254
	CH	141,261	12,719,203
	FR	61,777	6,397,943
	Total:	358,787	33,242,274
Dev	BE	7,723	824,871
	CA	171	17,061
	CH	5,244	476,338
	FR	4,864	434,547
	Total:	18,002	1,752,817
Test	BE	15,235	1,227,263
	CA	944	86,724
	CH	9,824	910,700
	FR	10,730	848,845
	Total:	36,733	3,073,532

Table 4: The FreCDo corpus is composed of about 400K data samples, containing a total of 38M tokens.

were identified using Spacy,[9] then replaced with the special token `$NE$`. Some statistics about the FreCDo corpus are presented in Table 4.

5.2 Participants and Approaches

Don't classify, translate (DCT): Instead of approaching dialect identification as a classification task, Gillin (2022) treated French variety identification as a translation task where the input text is the source, and the language labels are the target. To simplify the vocabulary used in the encoder-decoder model, the authors set FR, BE, CH, and CA as reserved symbols and allowed the vocabulary to be shared for both encoder and decoder. They employed a model inspired by Li et al. (2018), using slightly modified scripts from Susanto et al. (2019) to train the model. The DCT team submitted two closed runs with different architectures. The first run is based on an encoder with 6 layers, a decoder with 2 layers, and 8 attention heads. There are three models trained with different random seeds, which are combined into an ensemble. The second run is based on a similar ensemble, but the architecture is shallower, being formed of an encoder with 1 layer, a decoder with 1 layer, and 1 attention head. For cases in which the translation model fails (e.g. when returning blank labels), the authors fall back to the FR label.

NRC: The NRC team (Bernier-Colborne et al., 2022) submitted three closed runs and three open runs. They constructed a majority vote ensemble for the first closed run based on five multi-class SVMs trained on the joint training and development data, using different data processing and feature sets. The differences between the models involve the usage of word tokenization, the removal of redundant NE tokens, the filtering of training data using a minimum text length threshold, and the usage of n-grams as features. Three of the models used only word bigrams as features, while the other two used word unigrams and bigrams, as well as character trigrams and 4-grams. The authors carried out a greedy search among a dozen SVM models, looking at the results on the development set to select the best subset of models. For the second closed run, the authors employed a probabilistic classifier similar to Naïve Bayes, trained on the concatenation of the training and development data, as well as the pseudo-labeled test data, where the test labels are those predicted by the SVM ensemble used for their first run. The feature set used by this classifier includes only word bigrams. The third closed run is based on a single multi-class SVM classifier, providing the best development data results. This model was trained on the concatenation of the training and development data, using only word bigrams as features.

The open runs submitted by the NRC team are all based on variants of CamemBERT (Martin et al., 2020). The first open run is based on a majority vote ensemble of 3 pre-trained CamemBERT models, which were fine-tuned on the concatenation of the training and development data, starting with the pre-trained encoder weights and tokenizer. The authors performed model selection based on the scores obtained on the development data. The differences between the three models involve the batch size (8 or 16), the learning rate schedule (constant or linear decay), and the number of encoder layers that were fine-tuned (either just the last layer or the last two layers). For the second open run, the NRC team relied on their best single CamemBERT model according to the results on the development set, fine-tuned on the joint training and development data. This model was fine-tuned with a batch size of 8 and a constant learning rate for 3 epochs. Only the last 2 layers of the encoder were fine-tuned. For the third open run, the team employed

[9] https://spacy.io

Rank	Team	Run	Macro-F1	Micro-F1
1	NRC	2	0.3437	0.4936
2	NRC	1	0.3266	0.4642
3	NRC	3	0.3149	0.4530
4	SUKI	3	0.2661	0.3918
5	DCT	1	0.2627	0.3914
6	SUKI	1	0.2603	0.3984
7	DCT	2	0.1905	0.3421
8	SUKI	2	0.1383	0.2339

Table 5: F1-scores attained by the teams participating in the 2022 FDI **closed** shared task.

Rank	Team	Run	Macro-F1	Micro-F1
1	NRC	1	0.4299	0.5243
2	NRC	3	0.4145	0.4936
3	NRC	2	0.4108	0.5067
–	Organizers	Baseline	0.3967	0.5584

Table 6: F1-scores attained by the teams participating in the 2022 FDI **open** shared task.

their second-best single CamemBERT model. The last 2 layers of this model were fine-tuned using a batch size of 16 for 5 epochs with linear learning rate decay.

SUKI: Jauhiainen et al. (2022) employed a custom-coded language identifier using the product of relative frequencies of character n-grams. The model is essentially a Naïve Bayes classifier using the relative frequencies as probabilities, being inspired by Jauhiainen et al. (2019). The authors only applied pre-processing to replace number-characters with '1'. The length of the character n-grams is set to 8. Instead of multiplying the relative frequencies, the authors summed up their negative logarithms. As a smoothing value, they used the negative logarithm of an n-gram appearing only once multiplied by a penalty modifier. The penalty modifier is set to 1.26. In addition, the SUKI team used the same language model adaptation technique as in their previous work (Jauhiainen et al., 2018). The adaptation to the test data is performed for 3 epochs, following Jauhiainen et al. (2019). In the end, the system is identical to the one used to win the RDI shared task 2021 (Chakravarthi et al., 2021), with some slight differences in pre-processing only (Jauhiainen et al., 2021). The SUKI team submitted three runs. The first run is based on considering the training data as training material, the second run uses the development data as training material, and the third run takes both the training and development data as training material. All runs are closed.

Baseline: Găman et al. (2022) introduced a CamemBERT model as baseline for the FreCDo corpus. The text is first tokenized with the CamemBERT tokenizer, obtaining 768-dimensional embedding vectors. Each sequence is then represented as a Continuous Bag-of-Words (CBOW) via appending a global average pooling layer. The final predictions are given by a Softmax classification layer. The whole model is fine-tuned for 30 epochs on mini-batches of 32 samples, using the AdamW optimizer (Loshchilov and Hutter, 2019).

5.3 Results

Evaluation measure: With the release of the test set, the participants were announced that the macro-averaged F1-score would be used to rank the submitted runs. For completeness, we also report the micro-averaged F1-score (which is equivalent to accuracy).

Closed: Table 5 presents the results for the 2022 FDI closed shared task. The NRC team's probabilistic model achieves the best score, closely followed by the SVM ensemble that was used to convey pseudo-labels for the test set to the top scoring model. The NCR team's best single SVM model ranked third. Interestingly, the SUKI team also pro-

Language	Dialect	F1	Exact Match	# Dev Questions	# Test Questions
Arabic	Algeria (DZA)	71.72	56.17	324	921
	Egypt (EGY)	72.39	56.39	324	921
	Jordan (JOR)	73.27	57.41	324	921
	Tunisia (TUN)	73.55	57.71	324	921
	Avg.	71.72	56.17	*Total* 1296	3684
English	Australia (AUS)	73.67	59.52	494	440
	India-South (IND-S)	72.22	58.10	494	440
	Nigeria (NGA)	73.36	58.70	494	440
	Philippines (PHI)	73.76	59.11	494	440
	USA-Southeast (USA-SE)	74.35	59.31	494	440
	Avg.	73.47	58.95	*Total* 2470	2200
Kiswahili	Kenya (KEN)	72.12	63.1	1000	472
	Tanzania (TZN)	70.74	61.7	1000	463
	Avg.	72.60	59.71	*Total* 2000	935

Table 7: DialQA baseline results (development set) on Answer Selection task.

posed a probabilistic model, but their results seem considerably lower. The main difference between the two probabilistic models, the one submitted by NRC and the other submitted by SUKI, seems to be the use of word n-grams in favor of character n-grams. Although character n-grams have been found useful in dialect identification in other languages, e.g. Arabic (Ionescu and Butnaru, 2017) or Romanian (Găman and Ionescu, 2022; Jauhiainen et al., 2021), it appears that word n-grams are more discriminative for French dialect identification on FreCDo. Perhaps using an entire range of character n-grams would have been a better choice for the SUKI team than just character 8-grams. The model employed by DCT stands out due to its unusual approach based on translation. Unfortunately, applying a translation model to the dialect identification task did not seem to pay off for the DCT team. Their models landed in ranks five and seven.

Open: NRC was the only team to submit runs for the 2022 FDI open shared task. The corresponding results are shown in Table 6. Here, the ensemble of CamemBERT models yielded the top score, but the individual CamemBERT models (second and third runs) also attained very good results. Comparing the open runs with the closed ones, it becomes clear that pre-trained language models benefit a lot from the large-scale corpora used to train the respective models, even if pre-training is carried out in a self-supervised manner.

5.4 Summary

For the French Dialect Identification shared task, we proposed a cross-domain four-way classification task. We received a total of eight closed submissions and three open submissions coming from three different teams. Each team submitted between two and six runs. Considering the results of the shared task participants and those attained by the baseline proposed with the dataset (Găman et al., 2022), we conclude that the cross-domain four-way FDI task remains very challenging, leaving sufficient room for future exploration. Basic machine learning models, e.g., Naïve Bayes or SVM, attained the strongest results in the closed setting. In the open scenario, we observed that using pre-trained language models is beneficial.

6 Dialectal Extractive Question Answering (DialQA)

6.1 Dataset

The task builds on the existing QA benchmarks TyDi-QA (Clark et al., 2020) and SD-QA (Faisal et al., 2021): specifically, it uses portions of the SD-QA dataset, which recorded dialectal variations of TyDi-QA questions. The original SD-QA dataset includes more than 68k audio prompts in 24 dialects from 255 annotators. In DialQA, we include

development and test data for five varieties of English (Nigeria, USA, South India, Australia, Philippines), four varieties of Arabic (Algeria, Egypt, Jordan, Tunisia), and two varieties of Kiswahili (Kenya, Tanzania). The recorded and transcribed questions are highly parallel across the dialects within a language.

6.2 Approach and Baselines

Answer Selection Task: In the first part, we provide a text-based extractive QA baseline. Here, we fine-tune mBERT (Devlin et al., 2019) on a modified TyDi-QA training dataset so that, given the question and a single passage, the system returns the start and end byte indices of the minimal span that answers the question (Alberti et al., 2019). The baseline is prepared within the constraints of SQuAD (Rajpurkar et al., 2016) Question Answering settings. So all the unanswerable questions are discarded beforehand while preparing the DialQA dev and test set.

Automatic Speech Recognition (ASR) Task: This second part is an open task defined over the utterances of the different language varieties. Given the audio file of the utterance, the model has to produce an accurate transcription to be provided as input to the text-based QA system.

6.3 Discussion

Table 7 presents the baseline scores (development set) for the Answer Selection part. We calculate both dialect and language level F1 and exact match scores. The F1-score varies from 70.7 to 74.4, with USA-Southeast English being the best performing variety. The difference in performance can largely be attributed to the dialect-level differences induced as transcription noise. For the second task, no baseline is provided. However, the difference across dialectal audios and their corresponding transcription could be considered to design an ASR module. Another possibility could be designing an end-to-end speech-to-text extractive QA system capable of taking the dialectal audios as input.

6.4 Summary

We propose an extractive Dialectal Question Answering task that is open to both text and audio questions as the system input. Along with the task, we release the dialectal development and test datasets. The task is still open for submission and further development. The data and baselines are freely available on GitHub: `https://github.com/ffaisal93/DialQA`.

7 Conclusion

This paper presented an overview of the three shared tasks organized as part of the VarDial Evaluation Campaign 2022: Identification of Languages and Dialects of Italy (ITDI), French Cross-Domain Dialect Identification (FDI), and Dialect Extractive Question Answering (DialQA).

Participants of these shared tasks were provided with existing or new data sets made available to the community, which were discussed in detail in the respective sections. We furthermore included short descriptions of each team's systems, along with references to all system description papers published in the VarDial workshop proceedings (Table 1). We compared the participants' contributions with the organizer-provided baselines and found that participants were able to beat the latter both in ITDI and in the open FDI track.

For the ITDI task, we observed that shallow machine learning models outperformed deep learning models – even when using pre-trained language models for Italian. In contrast, pre-trained French language models provided much better performances than shallow models in the FDI task. It seems therefore that the optimal model choice for language and dialect identification tasks is largely task-dependent. This confirms the findings of previous editions of the VarDial campaign (Chakravarthi et al., 2021; Zampieri et al., 2020), where similar diverging trends were observed.

Acknowledgements

We thank all the participants for their interest in the shared tasks.

The work related to the ITDI shared task has received funding from the Swiss National Science Foundation (project no. 191934) and the Academy of Finland (project no. 1342859). The work related to the DialQA task was supported by the US National Science Foundation (project no. IIS-2125466).

References

Chris Alberti, Kenton Lee, and Michael Collins. 2019. A BERT baseline for the natural questions. *arXiv preprint arXiv:1901.08634*.

Antonio Ciccolella via Wikimedia Commons. 2015. Italiano: Mappa delle lingue e gruppi dialettali italiani. https://upload.wikimedia.org/wikipedia/commons/3/32/Dialetti_e_lingue_in_Italia.png.

Giuseppe Attardi. 2015. WikiExtractor. https://github.com/attardi/wikiextractor.

Gabriel Bernier-Colborne, Serge Leger, and Cyril Goutte. 2022. Transfer Learning Improves French Cross-Domain Dialect Identification: NRC @ VarDial 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, Gyeongju, Republic of Korea. International Committee on Computational Linguistics (ICCL).

Giacomo Camposampiero, Quynh Anh Nguyen, and Francesco Di Stefano. 2022. The curious case of Logistic Regression for Italian Dialect Identification: ETHZ team at ITDI Vardial 2022 Shared Task. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, Gyeongju, Republic of Korea. International Committee on Computational Linguistics (ICCL).

Andrea Ceolin. 2022. Comparing the performance of CNNs and shallow models for language identification. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, Gyeongju, Republic of Korea. International Committee on Computational Linguistics (ICCL).

Bharathi Raja Chakravarthi, Gaman Mihaela, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial evaluation campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–11, Kyiv, Ukraine. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. Ethnologue: Languages of the world. twenty-fifth edition. http://www.ethnologue.com/. Dallas, Texas: SIL International.

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. SD-QA: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Nat Gillin. 2022. Is Encoder-Decoder Transformer the Shiny Hammer? In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, Gyeongju, Republic of Korea. International Committee on Computational Linguistics (ICCL).

Mihaela Găman, Adrian Gabriel Chifu, William Domingues, and Radu Tudor Ionescu. 2022. FreCDo: A New Corpus for Large-Scale French Cross-Domain Dialect Identification. *(under review)*.

Mihaela Găman and Radu Tudor Ionescu. 2022. The Unreasonable Effectiveness of Machine Learning in Moldavian versus Romanian Dialect Identification. *International Journal of Intelligent Systems*, 37(8):4928–4966.

Radu Tudor Ionescu and Andrei Butnaru. 2017. Learning to identify arabic and german dialects using multiple kernels. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 200–209, Valencia, Spain.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2018. HeLI-based experiments in Swiss German dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 254–262, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2021. Naive Bayes-based experiments in Romanian dialect identification. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 76–83, Kyiv, Ukraine. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022. Italian Language and Dialect Identification and Regional French Variety Detection using Adaptive Naive Bayes. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, Gyeongju, Republic of Korea. International Committee on Computational Linguistics (ICCL).

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2019. Discriminating between Mandarin Chinese and Swiss-German varieties using adaptive language models. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 178–187, Ann Arbor, Michigan. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Maggie Yundi Li, Stanley Kok, and Liling Tan. 2018. Don't classify, translate: Multi-level e-commerce product categorization via machine translation. *arXiv preprint arXiv:1812.05774*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of ICLR*.

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 the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 1–14, Osaka, Japan. The COLING 2016 Organizing Committee.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Alan Ramponi. 2022. NLP for language varieties of Italy: Challenges and the path forward. *arXiv preprint arXiv:2209.09757*.

Raymond Hendy Susanto, Ohnmar Htun, and Liling Tan. 2019. Sarah's participation in WAT 2019. In *Proceedings of the 6th Workshop on Asian Translation*, pages 152–158, Hong Kong, China. Association for Computational Linguistics.

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 the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

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 the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 1–17, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru, and Tommi Jauhiainen. 2019. A report on the third VarDial evaluation campaign. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–16, Ann Arbor, Michigan. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

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

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 1–9, Hissar, Bulgaria. Association for Computational Linguistics.

A Appendix

A.1 ITDI Data Sources

The development and test data come from the websites given below.

EML	https://www.bulgnais.com/libri.html
FUR	https://wikisource.org/wiki/Main_Page/Furlan
FUR	https://arlef.it/it/materiali
FUR	https://www.filologicafriulana.it/lenghe-e-culture
LLD	https://wikisource.org/wiki/Main_Page/Ladin
LLD	https://it.wikisource.org/wiki/Biancognee
LLD	https://www.istitutoladino.it
LIJ	https://lij.wikisource.org
LMO	https://wikisource.org/wiki/Main_Page/Lumbaart
LMO	https://www.lingualombarda.it/index.php/milanese.html
NAP	https://nap.wikisource.org
NAP	https://it.wikisource.org/wiki/Categoria:Testi_in_napoletano
PMS	https://pms.wikisource.org
SC	https://wikisource.org/wiki/Category:Sardu
SCN	https://wikisource.org/wiki/Category:Sicilianu
SCN	https://wikisource.org/wiki/Main_Page/Sicilianu
SCN	https://it.wikisource.org/wiki/Categoria:Testi_in_siciliano
SCN	http://www.linguasiciliana.org
SCN	http://www.salviamoilsiciliano.com/raccolte
SCN	http://www.museomirabilesicilia.it/folklore-siciliano.html
SCN	http://www.salviamoilsiciliano.com/raccolte
SCN	http://rapallosalvatore.blogspot.com/p/raccolta-poesie-in-dialetto-siciliano.html
ROA-TARA	http://www.tarantonostra.com
VEC	https://vec.wikisource.org
several	https://www.dialettando.com