A Report on the Third VarDial Evaluation Campaign

Marcos Zampieri[1], Shervin Malmasi[2], Yves Scherrer[3], Tanja Samardžić[4]
Francis Tyers[5], Miikka Silfverberg[3], Natalia Klyueva[6], Tung-Le Pan[6]
Chu-Ren Huang[6], Radu Tudor Ionescu[7], Andrei Butnaru[7], Tommi Jauhiainen[3]

[1]University of Wolverhampton, [2]Harvard Medical School, [3]University of Helsinki
[4]University of Zurich, [5]Indiana University, [6]The Hong Kong Polytechnic University
[7]University of Bucharest

m.zampieri@wlv.ac.uk

Abstract

In this paper, we present the findings of the Third VarDial Evaluation Campaign organized as part of the sixth edition of the workshop on Natural Language Processing (NLP) for Similar Languages, Varieties and Dialects (VarDial), co-located with NAACL 2019. This year, the campaign included five shared tasks, including one task re-run – German Dialect Identification (GDI) – and four new tasks – Cross-lingual Morphological Analysis (CMA), Discriminating between Mainland and Taiwan variation of Mandarin Chinese (DMT), Moldavian vs. Romanian Cross-dialect Topic identification (MRC), and Cuneiform Language Identification (CLI). A total of 22 teams submitted runs across the five shared tasks. After the end of the competition, we received 14 system description papers, which are published in the VarDial workshop proceedings and referred to in this report.

1 Introduction

The series of workshops on Natural Language Processing (NLP) for Similar Languages, Varieties and Dialects (VarDial) has reached its sixth edition in 2019, evidencing the interest of the CL/NLP community in this topic. The third VarDial Evaluation Campaign[1] featuring five shared tasks, described in detail in this report, has been organized as part of VarDial 2019 co-located with the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). It follows two editions of the campaign organized in 2017 with four tasks (Zampieri et al., 2017) and in 2018 with five tasks (Zampieri et al., 2018).

Since its first edition, shared tasks have been organized as part of the VarDial, most notably the Discriminating between Similar Languages organized from 2014 to 2017 (Zampieri et al., 2014, 2015; Malmasi et al., 2016). The shared tasks organized at VarDial helped providing evaluation benchmarks and public datasets (e.g. (Tan et al., 2014)) for different tasks such as dialect identification, morphosyntactic tagging, and cross-lingual dependency parsing. Similar languages such as Bulgarian and Macedonian, and Czech and Slovak, along with varieties and dialects of Arabic, German, Hindi, Portuguese, and Spanish have been included in the competitions organized within the scope of VarDial.

In this paper, we present the results and main findings of the third VarDial Evaluation Campaign. The five tasks organized this year were: German Dialect Identification (GDI) presented in Section 4, Cross-lingual Morphological Analysis (CMA) presented in Section 5, Discriminating between Mainland and Taiwan variation of Mandarin Chinese (DMT) presented in Section 6, Moldavian vs. Romanian Cross-dialect Topic identification (MRC) presented in Section 7, and finally, Cuneiform Language Identification (CLI) presented in Section 8. In Table 1, we include references to the 14 system description papers written by the participants of the campaign and published in the VarDial workshop proceedings.

2 Shared Tasks at VarDial 2019

The five shared tasks organized as part of the VarDial Evaluation Campaign 2019 are listed next:

Third German Dialect Identification (GDI): After two successful editions of the (Swiss) German Dialect Identification task, we organized a third iteration of this task at VarDial 2019. We focused again on four Swiss German dialect

[1] https://sites.google.com/view/vardial2019/campaign

Proceedings of VarDial, pages 1–16
Minneapolis, MN, June 7, 2019

areas (Basel, Bern, Lucerne, and Zurich). We provided updated speech transcripts for all dialect areas, but also released two complementary data sources: acoustic data in the form of iVectors, and (predicted) word-level normalisation. In particular, the Arabic Dialect Identification (ADI) task organized in previous VarDial evaluation campaigns showed that acoustic features may substantially improve dialect identification. We wanted to investigate whether this also holds in the slightly different GDI setting.

Cross-lingual Morphological Analysis (CMA): At VarDial 2019, we introduce the task of cross-lingual morphological analysis. Given a word in an unknown related language, for example "navifraghju" ("shipwreck" in Corsican), a human speaker of several related languages is able to deduce that it is a noun in the singular by making deductions from similar words, for example: "naufrag" (Catalan), "naufragio" (Spanish, Italian), "naufrágio" (Portuguese), "naufrage" (French) and "naufragiu" (Romanian). At CMA, we invited participants to create computational models able to do the same. Two language families were represented in the dataset, Romance (fusional morphology) and Turkic (agglutinative morphology). In the "Closed" track, participants were given a set of word forms with all valid morphological analyses in six languages and asked to predict the valid morphological analyses for a seventh, unseen language. In the "Semi-Closed" track, the process was the same, only participants were provided with additional raw data by the organisers. This was in the form of raw text Wikipedia dumps, bilingual dictionaries from the Apertium project and any treebanks available in the known languages from the Universal Dependencies project.

Discriminating between Mainland and Taiwan variation of Mandarin Chinese (DMT): Like English, Mandarin has several varieties among the speaking communities and two dominant standard varieties (Lin et al., 2018). This task aims to discriminate between these two standard varieties of Mandarin Chinese: Putonghua (Mainland China) and Guoyu (Taiwan). We provide a corpus of approximately 10,000 sentences from newspapers for each Mandarin variety. The main task is to determine if a sentence is written in the Mandarin variety of Mainland China or from Taiwan. It is important to note that since a direct consequence and the most salient feature of the variations is the use of different orthographic systems in China (simplified) and Taiwan (traditional), so the task is designed to focus on the linguistic rather than orthographic differences. Each sentence in the corpus is tokenized and punctuations are removed from the texts, as well as converted from original traditional orthography to simplified, and vice versa. Hence both the traditional and the simplified versions of the same corpus are available so that participant can choose either version and won't be able to use orthographic cues. The results are evaluated in two separate tracks (Simplified and Traditional).

Moldavian vs. Romanian Cross-dialect Topic identification (MRC): In the Moldavian vs. Romanian Cross-topic Identification shared task, we provided participants with the MOROCO data set (Butnaru and Ionescu, 2019) which contains Moldavian and Romanian samples of text collected from the news domain. The samples belong to one of the following six topics: culture, finance, politics, science, sports, and tech. The samples are pre-processed in order to eliminate named entities. For each sample, the data set provides corresponding dialectal and category labels. To this end, we proposed three subtasks for the 2019 VarDial Evaluation Campaign. The first sub-task was a binary classification by dialect task, in which a classification model is required to discriminate between the Moldavian and the Romanian dialects. The second subtask was a Moldavian to Romanian cross-dialect multi-class classification by topic task, in which a model is required to classify the samples written in the Romanian dialect into six topics, using samples written in the Moldavian dialect for training. Finally, the third subtask was a Romanian to Moldavian cross-dialect multi-class classification by topic task, in which a model is required to classify the samples written in the Moldavian dialect into six topics, using samples written in the Romanian dialect for training.

Cuneiform Language Identification (CLI): This shared task focused on discriminating between languages and dialects originally written using the cuneiform script. The task included 2 dif-

Team	GDI	CMA	DMT	MRC	CLI	System Description Papers
Adaptcenter			✓			
BAM	✓					(Butnaru, 2019)
dkosmajac	✓					
DTeam				✓		(Tudoreanu, 2019)
SharifCL					✓	(Doostmohammadi and Nassajian, 2019)
ghpaetzold	✓		✓		✓	
gretelliz92			✓			
ekh					✓	
IUCL			✓			(Hu et al., 2019)
HSE		✓				(Mikhailov et al., 2019)
itsalexyang			✓			(Yang and Xiang, 2019)
lonewolf				✓		
MineriaUNAM		✓				
NRC-CNRC					✓	(Bernier-Colborne et al., 2019)
R2I_LIS				✓		(Chifu, 2019)
PZ					✓	(Paetzold and Zampieri, 2019)
SC-UPB				✓		(Onose and Cercel, 2019)
situx					✓	
SUKI	✓		✓			(Jauhiainen et al., 2019b)
tearsofjoy	✓		✓	✓	✓	(Wu et al., 2019)
TübingenOslo		✓				(Çöltekin and Barnes, 2019)
Twist Bytes	✓				✓	(Benites et al., 2019)
Total	**6**	**3**	**7**	**5**	**8**	**14**

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

ferent languages: Sumerian and Akkadian. Furthermore, the Akkadian language was divided into six dialects: Old Babylonian, Middle Babylonian peripheral, Standard Babylonian, Neo Babylonian, Late Babylonian, and Neo Assyrian. These languages and dialects were used in ancient Mesopotamia and span a time period of 3,000 years. For training and development, we provided the participants with varying amounts of text encoded in Unicode cuneiform signs for each language or dialect.

3 Participating Teams

The Third VarDial Evaluation Campaign received a positive response from the NLP community. A total of 51 teams enrolled to participate in the five shared tasks of the campaign and 22 of them submitted runs to one or more tasks. This is a similar participation rate to VarDial 2018 when 54 teams signed up and 24 teams submitted runs to five shared tasks, a record for the workshop.

In VarDial 2019, the participants could choose to participate in one or more shared tasks. Table 1 lists the participating teams, the shared tasks they took part in, and a reference to each of the 14 system description papers published in the VarDial workshop proceedings.

4 Third German Dialect Identification GDI

The third edition of the (Swiss) German Dialect Identification task was based on the same data source and split as in 2018, but offered the participants the possibility to make use of word-level normalizations and/or acoustic features. The GDI task again covered four Swiss German dialect areas, namely Basel, Bern, Lucerne, and Zurich.

4.1 Dataset

As in 2017 and 2018, we extracted the training and the test datasets from the ArchiMob corpus of Spoken Swiss German (Samardžić et al., 2016; Scherrer et al., in press). This corpus currently contains 43 oral history interviews with informants speaking different Swiss German dialects. Each interview was transcribed by one of four transcribers, using transcription guidelines based on the writing system "Schwyzertütschi

Dialäktschrift“ (Dieth, 1986). The transcriptions exclusively used lower case.

We provided the same data splits as in 2018, but with slightly reduced sizes due to additional filtering. The training set contained utterances from at least three interviews per dialect. The development and test sets each contained utterances from at least one other interview per dialect. Participants were encouraged to include the development data as additional training material in their final systems. This year, we also provided word-level normalizations and acoustic features.

The normalizations have been produced automatically using character-level statistical machine translation at utterance level and re-aligning the normalizations with their source words (see Scherrer and Ljubešić (2016) for details on the approach). We estimated that this word-level normalization format would allow participants to experiment with various feature representations such as character alignments. The normalization language resembles Standard German, but deviates from it in many respects.

The acoustic features, in the form of 400-dimensional i-vectors, were extracted from the source audio data, aligned with the text at the level of segments whose length is between 4s and 10s. Our extraction procedure follows closely the steps proposed in the previous work on Arabic dialects (Ali et al., 2016; Dehak et al., 2011). As in the previous work, we use the Kaldi collection of tools[2] to perform different calculations needed for the extraction of i-vectors. While i-vectors are expected to model the difference between individual speakers and the general background model, the question is open whether they offer some reliable dialect-level information, which can be exploited by the classification algorithms. Given that there is no speaker overlap between training and test data in our current GDI setup, dialect-level information is necessary for improving over the baseline.

4.2 Participants and Approaches

Six participants submitted their systems to the GDI task this year. In the following paragraphs, we shortly describe the best system submitted by each participant. Many participants also provided alternative systems.

[2] https://github.com/kaldi-asr/kaldi/blob/08869e31da51d688ee582dc924193b19530a2d32/egs/lre07/v1/lid/extract_ivectors.sh

tearsofjoy: This submission is based on a linear SVM classifier using character 1–5-grams, word 1–2-grams as well as the iVector features. The character and word features are weighted by BM25. Semi-supervised adaptation to the test data was also used.

SUKI: This submission uses the HeLI method, which is based on relative frequencies of character 4-gram features with smoothing. One of its key characteristics is the semi-supervised adaptation to the test data, as proposed in 2018.

Twist Bytes: This submission relies on a SVM meta-classifier that uses multiple tf-idf-weighted character and word features. Acoustic features are used in a base SVM classifier, whose predictions serve as input for the meta-classifier. Semi-supervised adaptation to the test data was also used.

BAM: This system is an ensemble of three models, a character-level convolutional neural network, a character-level LSTM, and a string kernel model.

dkosmajac: This submission relies on a quadratic discriminant analysis classifier for the iVectors and on a random forest classifier for the text. The output of both classifiers is fed into a random forest meta-classifier to produce the final predictions.

ghpaetzold: This system consists of a recurrent neural network that learns representations of sentences based on their words, and of words based on their characters.

The baseline consists of a linear SVM classifier using only word unigrams as features.

4.3 Results

Table 2 shows the performance of different methods on the GDI data in terms of macro-averaged F1 scores. The three best models all include semi-supervised adaptation to the test data. The impact of the iVectors is hard to assess: on the one hand, it was expected to be low due to the lack of speaker overlap between training and test data, but on the other hand semi-supervised adaptation should be able to generalize test speaker properties from the acoustic signal. The results do not bear out this second hypothesis. None of the participants used

Rank	Team	Transcripts	iVectors	Normalization	Adaptation	F1 (macro)
1	tearsofjoy	✓	✓		✓	0.7593
2	SUKI	✓			✓	0.7541
3	Twist Bytes	✓	✓		✓	0.7455
4	BAM	✓				0.6255
	Baseline	✓				*0.6078*
5	dkosmajac	✓	✓			0.5616
6	ghpaetzold	✓				0.5575

Table 2: Results and rankings of GDI participants. The table also specifies the data formats and techniques used by the participants.

the normalized data. As in previous years, systems based on neural networks did not reach competitive scores, possibly also due to the absence of adaptation.

4.4 Summary

In this third iteration of the GDI task, we provided additional data formats such as acoustic data and word-level normalizations. Six teams participated in the GDI task. Three of them used the acoustic data, but results do not seem to indicate large gains. In contrast, semi-supervised adaptation to the test set seems to be crucial to attain state-of-the-art results.

5 Cross-lingual Morphological Analysis (CMA)

Morphological analysis is one of the cornerstones of natural language processing for morphologically complex languages. Currently, rule-based finite-state morphological analyzers represent the state-of-the-art for this task, however, developing rule-based analyzers is a substantial task. It entails creation of extensive word lists and grammatical descriptions. This requires both linguistic expertise and technical expertise in the rule formalism which is used. Hence, there exists a demand for less labor intensive approaches especially for low-resource languages.

Classically, rule-based analyzers have been augmented with statistical guessers which provide analyses for out-of-lexicon word forms (Lindén, 2009). Recently, purely data-driven morphological analysis has received increasing attention (Nicolai and Kondrak, 2017; Silfverberg and Hulden, 2018; Moeller et al., 2018; Silfverberg and Tyers, 2019). Purely data-driven systems learn an analysis model from a data set of morphologically analyzed word forms and can then be applied to unseen word forms.

The shared task on cross-lingual morphological analysis (CMA) investigates a new dimension of the morphological analysis task. The task was to leverage data for related languages in building a purely data-driven analyzer for a target language. No annotated target language data was provided to the competitors.

The CMA task investigated related-language analysis for the Romance and Turkic language families. Competitors were provided morphologically analyzed training data in six Romance languages (Asturian, Catalan, French, Italian, Portuguese and Spanish) and six Turkic languages (Bashkir, Crimean Tatar, Kazakh, Kyrgyz, Tatar and Turkish). Using these datasets, they built morphological analyzers for two surprise languages: the Romance language Sardinian and the Turkic language Karachay-Balkar. The competitors had access to the input word forms in the Sardinian and Karachay-Balkar test sets but, as stated above, they did not receive any morphologically analyzed data in either of the target languages.

5.1 Dataset

The dataset was compiled specifically for the shared task. We used the Wikipedias in all the languages to create a frequency list of surface tokens for each language. We then analysed these lists using the morphological analysers from the Apertium (Forcada et al., 2011) project. The lists of analyses were trimmed to include only open-class parts of speech (nouns, adjectives, adverbs and verbs). We then removed any form which did not include at least one analysis in an open class. After this we took the top 10,000 wordforms for each language.

The tagsets were converted from Apertium-style to Universal Dependencies (Nivre et al.,

Team	Turkic			Romance		
	Analysis	Lemma	Tag	Analysis	Lemma	Tag
HSE	35.61	**56.99**	38.75	*23.28*	*38.82*	*46.42*
MineriaUNAM	0.00	0.56	0.00	0.33	0.44	37.76
TübingenOslo	31.53	52.74	38.93	23.67	31.36	**61.33**
BASELINE-I	**39.46**	54.94	44.18	22.94	31.56	51.88
BASELINE-II	39.44	53.82	**44.29**	**26.51**	**34.65**	58.54

Table 3: Results for the CMA task. Bold indicates the best scoring system, while *italics* indicates an 'unofficial' result that was submitted after the deadline. These scores are F-scores. For the Analysis column every part of the analysis had to be correct, for the Lemma column the lemma had to be correct and for the Tag column just the part-of-speech tag had to be correct. BASELINE-I refers to the neural system and BASELINE-II to the neural ensemble described in Section 5.3.

2016) using a longest-match set overlap method running on tag-lookup tables, for example, the Apertium tag `<n>` was converted to the Universal Dependencies tag `NOUN`, while Apertium's `<p1>` was translated into Universal Dependencies `Number=Plur|Person=1`.

Finally, each of the word forms was labelled with the language it came from and the lists were merged into language family specific lists.

5.2 Participants and Approaches

HSE This team constructed a POS specific cross-lingual morpheme inventory using the annotated training data. They then predicted target language POS tags using a bidirectional LSTM encoder-decoder model with attention. Finally, they used the POS specific morpheme inventory to predict morphological features using a greedy algorithm. Lemmatization was accomplished by suffix stripping. To deal with language specific orthographic conventions, the team first automatically transcribed all the training data into a joint orthographic representation: For Romance languages, diacritics were removed and for Turkic languages, all data sets were transcribed into Cyrillic script. To build the morpheme inventories, word forms were morphologically segmented using Morfessor (Smit et al., 2014).

MinerialUNAM No system description paper was submitted by this team.

TübingenOslo This team divided the morphological analysis task into two sub-tasks: lemmatization and morphological tag prediction. First, a bidirectional GRU encoder was used to encode the input word form into a representation vector. This vector was fed into a GRU decoder network which generated a lemma. A number of feed forward networks were then used to predict morphological features and POS tag using the representation vector as input. Each morphological feature type, for example number and case, was predicted by a separate feed forward network. Additionally, this team reports results for a linear baseline system which delivers competitive performance for the Turkic language family.

5.3 Baseline System

The first baseline system BASELINE-I (Silfverberg and Tyers, 2019) formulates the morphological analysis task as a character-level string transduction task. It uses an LSTM encoder-decoder model with attention (Bahdanau et al., 2014) for performing the string transduction. To this end, the system is trained to translate input word forms like *andaluza* (feminine singular for the noun or adjective *andaluz* 'Andalusian' in Spanish) into a set of output analyses: `andaluz+A+Num=Sg|Gend=Fem` and `andaluz+N+Number=Sg|Gend=Fem`.

Since a word form may have multiple valid morphological analyses with different lemmas, POS tags and MSDs (for example, *andaluza* has two), the baseline model needs to be able to generate multiple output analyses given an input word form. This is accomplished by extracting several output candidates from the model using beam search and selecting the most probable candidates as model outputs. The number of outputs is controlled by a probability threshold hyperparameter p. The system extracts the least number of top scoring candidates whose combined probability mass is greater than p. Additionally, the number of output candidates is restricted using a single

hyperparameter N which is a firm upper bound for the number of analyses a word may receive. The hyperparameters p and N are tuned by treating the training set for one of the languages as held-out data (Asturian for Romance languages and Crimean Tatar for Turkic languages). After tuning the hyperparameters, the model was trained on the complete annotated training data.

The second baseline system BASELINE-II is an ensemble of five instances of the neural baseline systems BASELINE-I described above. Each instance was trained identically apart from random initialization of model parameters. We compute the probability for an output analysis as the arithmetic mean of the probabilities assigned by each of the five component models. Output analyses are generated in the same manner as for the BASELINE-I model.

5.4 Results

Given an input word form, systems return a set of analyses each of which consists of a lemma and a morphological tag. Systems are evaluated for F1-score with regard to the gold standard set of complete analyses, lemmas and tags for each input word form. Table 3 shows results for the CMA task.

5.5 Summary

Three teams participated in this first iteration of the cross-lingual analysis task. Two of the teams employed variations of neural encoder-decoder systems. Apart from lemmatization performance, it proved to be difficult to attain consistent improvements over the neural baseline systems. However, the suffix stripping approach used by the HSE team did deliver clear improvements in lemmatization for both Turkic and Romance languages.

6 Discriminating between Mainland and Taiwan variation of Mandarin Chinese (DMT)

Mandarin, with over 900 million native speakers, is one of the ten main dialect groups of Chinese, along with Yue, Min, Wu, and others (often also referred to as Sinitic languages). Inside Mandarin, there is also a variety of divergence within. Mandarin (i.e. the language of the mandarins (the officers)) has been the official language of the government by convention for over a thousand years but has also become the common language both in spoken language and written text by constitution in the modern era, first by the Nationalists (ROC) after 1911, and then by the Communists PRC in 1949. In daily non-technical usage, both Chinese or Mandarin refers to either or both of these standard forms of Mandarin as the lingua franca of the Chinese people, including both their spoken and written forms (Huang and Shi, 2016). Although the later version (called Putonghua (普通话, common language) superseded the older version (called Guoyu (國語, national language) in Mainland China, and the latter version persists in Taiwan and can be viewed to be related, important variations arose since 1949 for several reasons (Lin et al., 2018).

First, and most of all, the two varieties developed in relative isolation from each other and under different political systems for over 50 years during the Cold War era. Second, each has its own regulating bodies as well as different contextual influences. Third, Guoyu has more southern influences than Putonghua, even though both are based on Beijing Mandarin. Note that Putonghua in China is written with simplified Chinese characters with Pinyin romanization for pedagogy; while Guoyu in Taiwan is written in traditional characters and uses the Zhuyin system (sometimes called bopomofo) for pedagogy. With recent more frequent exchanges at different levels of China and Taiwan, some of the differences have begun to get absorbed.

6.1 Dataset

Texts to distinguish between the two variations were compiled from the two existing corpora of news: Sinica Corpus for Taiwan Mandarin (Chen et al., 1996) and LCMC (The Lancaster Corpus of Mandarin Chinese, (McEnery and Xiao, 2003)) for Mainland Mandarin. Both corpora are segmented and tokenized. We remove the punctuation and unify the orthography used to eliminate orthographic cues. Since both corpora are balanced corpora, our initial thought was to provide genre-aware classification. However, inspection of both corpora suggested the genres were not defined in the same way and are not distributed homogeneously. In the next edition this idea may be exploited by using some additional resources as genre vs. regional variations which is an important and yet under-explored issue in similar languages

(Hou and Huang, 2019).

Thus, as input data, we got 21492 lines/sentences of LCMC corpus and 46158 lines/sentences from Sinica Corpus. The clean-up included removing lines containing Latin characters in Named Entities (as potential contextual cues) and lines shorter than 4 tokens. The LCMC portion is reduced to 12072 sentences after clean-up, and Sinica Corpus data is reduced correspondingly for balance.

The data were converted into utf-8 encoding, and split into training, development and test sets in the following proportions respectively for each variety: 9385/1000/1000 lines. Each of the sets was mixed pairwise: Taiwanese with Mainland train/dev/test sets, and shuffled. The test set was formed from the last 1000 lines of each of the corpus to make sure there is no intersection between training and test data.

The sets prepared as described above were then run through a character converter to form two tracks: Traditional and Simplified. As it was stated in the introduction of this Section, Mainland uses simplified characters while in Taiwan traditional characters are used. The conversion ensures that the DMT task is not orthography dependent and will allow us to compare results of teams working on both sets of data. Conversion from simplified to traditional and from traditional to simplified characters were made by opencc converter [3] (in effect, coding sets with some lexical conversion as well). However, conversion cannot be 100% accurate in both directions, it will have some information lost.

6.2 Participants and Approaches

A total of seven systems participated in the shared task, and as a result, 17 runs each were performed for both the simplified and the traditional set. Five of the teams performed three runs each for both sets of data and the other two only performed once for each. The results were given out in confusion matrices, which calculate the number of sentences that were identified as being labeled correctly and incorrectly. Four of the teams that participated in the shared task used the training and development data exclusively in order to obtain the final result.

Here is a more detailed explanation of the approaches conducted by the teams based on the descriptions provided by the participants:

Adaptcenter: A dictionary was built which contain the 5,000 high frequency words which were assigned values. Then the convolutional neural network (CNN) method was employed to the training and test test, which results in the CNN model. At the end, the two methods were combined, improving from either of the methods.

ghpaetzold: This system is a 2-layer compositional recurrent neural network that learns numerical representations of sentences based on their words, which were in turn based on their characters. The system receives, as input, the text from the instance being classified only, with no other additional features or resources. The model was trained exclusively on the training data provided, and was validated on the development set provided. The model was implemented in Pytorch.

gretelliz92: A simple preprocessing was carried out to preserve all the characteristics that can be discriminative between the two types of texts that are analyzed, with the combination of a linguistic feature based on tf-idf. Therefore, in this step only the texts with fasttext word embedding for Chinese were represented. The vectors obtained in the preprocessing are used as input of the model which consists of a Bidirectional long short-term memory (Bi-LSTM) layer, whose output is inputted to a fully connected neural network of 4 dense layers with the relu activation function, along with one output layer with the softmax function.

IUCL (submitted as 'hezhou'): An ensemble model was used containing the five following classifiers: 1) a pre-trained BERT model for Chinese, 2) a long short-term memory model with word-embeddings which was trained on People Daily's News, one of China's leading newspapers, 3) support vector machine (SVM) and Naive Bayes classifiers with word *n*-gram and context-free grammar features, 4) a sequential model with a global average pooling layer, 5) a word-based bi-LSTM model. They were, in turn, ensembled in three different methods: 1) assigning the class which has the highest probability (confidence) from any classifier, 2) assigning the class with the highest average probability, 3) using an SVM to predict the class from the probabilities given by all of the classifiers.

itsalexyang: A multinomial Naive Bayes and BiLSTM ensemble model was used to train the model. For multinomial Naive Bayes, it is trained using presence vs. absence (0 vs. 1) vectors based on feature combinations of character-level

[3] https://github.com/BYVoid/OpenCC

bigrams and trigrams as input. For BiLSTM, the Word2vec method was trained on the dataset to obtain word embedding matrices, then the word embedding sequences can be used as input sentence representations. A forward and a backward LSTM is used to process the sequence and produce hidden states, which contain information from contexts in two opposite directions. After obtaining the hidden state sequence, max-over-time pooling operation is applied to form a fix-size vector as sentence representations, which will be fed into a hidden dense layer with 256 units and a final dense layer to predict. After training with these base classifiers, an average of output probabilities from all the models is then taken and used to make the final prediction.

SUKI: A custom coded language identifier was made using the product of relative frequencies of character *n*-grams. It is a Naive Bayes classifier that uses relative frequencies as probabilities. The lengths of the character *n*-grams used ranged from 1 to 14 for the Traditional track and from 1 to 15 for the Simplified Track. Instead of multiplying the relative frequencies, their negative logarithms were summed up. As a smoothing value, the negative logarithm of an *n*-gram appearing only once multiplied by a penalty modifier was used. In this case, the penalty modifier was 1.3. For the Simplified track, similar language model (LM) adaptation was used as in GDI 2018 (Jauhiainen et al., 2018a). In addition, a separate confidence threshold was used. For the Traditional track, the LM adaptation was also used, but the results were split in 4 parts and all the information from one part was added to the language models at once. The *n*-gram models used, penalty modifier, the confidence threshold, and the number of splits in adaptation was optimized using the development data.

tearsofjoy: This is a linear SVM classifier (one-vs-rest multi-class classifier) with character *n*-grams ranging from the order 1 to 4 combined with word unigrams (as the effect of word *n*-grams on the development set is negligible). All *n*-gram features are combined into a single feature matrix and weighted by BM25. The model is tuned for optimum 'C' parameter (5.8 for this approach) and maximum *n*-gram order on the training/development set. The data was modified by adding the test instances that are classified with a classifier trained on the training set with high confidence to the training set, and re-training the classifier with the additional 'silver' data from the test set.

6.3 Results

In the Table 4 we present the results of the teams in terms of F1-scores alongside with the summary of the methods that they have employed in order to train a model. One of the teams (IUCL, marked in italics in the table) used additional resources (pre-trained word embeddings) while training.

6.4 Summary

From the obtained results we can see that sophisticated approaches involving Deep Learning models do not necessarily outperform the traditional methods like Naive Bayes or SVM. We have manually analysed the sentences that got wrong prediction for most systems. Majority of those sentences were of the generic themes, which suggests the key factor for identifying the variation was topical rather than grammatical.

Another observation coming from the confusion matrices: for some systems the percentage of cases when Mainland label was predicted while Taiwanese was the True label, sometimes was half as much than for the other way round.

Finally, comparing results from both tracks by the same team, it is shown that differences in F1 are general quite small and performance ranking is relatively stable and independent of the track (i.e. orthography). This reassures robustness of the set. It is interesting to observe though that the better performing teams tend to have bigger deviation than the teams with lower performance. For instance, the smallest delta (0.001249321) came from gretelliz92; while the higher delta (0.035017912) came from SUKI. While SUKI's performance is more than 15% higher (Delta F1 roughly 0.15).

While the default hypothesis was that the more robust system should be the one least affected by choice of orthographic representation, the DMT task results suggest that it would be the other way around. That is, the system that performs better in differentiating varieties of similar languages should be 'biased' to pick up the differences hence could be affected by representational variations. The least biased system (i.e. seemingly 'robust') in fact has the less discriminating power.

DMT - Traditional

Rank	Team	Method	Features used	F1
1	SUKI	Naive Bayes	ch. *n*-grams	0.9084
2	*IUCL*	*ens:BERT, LSTM, SVM etc.*	*word emb.*	*0.9008*
3	tearsofjoy	linear SVM	ch. and word *n*-grams	0.8843
4	itsalexyang	ens: Naive Bayes,BiLSTM	word2vec	0.8687
5	Adaptcenter	CNN	freq-based value assign.	0.8317
6	ghpaetzold	RNN	ch. numeral representations	0.7959
7	gretelliz92	bi-LSTM , Relu	tf-idf	0.7483

DMT - Simplified

Rank	Team	Method	Features used	F1
1	*IUCL*	*ens:BERT, LSTM, SVM etc.*	*word emb.*	*0.8929*
2	tearsofjoy	linear SVM	ch. and word *n*-grams	0.8737
3	SUKI	Naive Bayes	ch. *n*-grams	0.8734
4	itsalexyang	ens:Naive Bayes, BiLSTM	word2vec	0.8530
5	Adaptcenter	CNN	freq-based value assign.	0.8124
6	ghpaetzold	RNN	ch. numeral representations	0.7934
7	gretelliz92	bi-LSTM, Relu	FastText word embeddings	0.7496

Table 4: The macro F1-scores for DMT-Traditional and DMT-Simplified shared task alongside with the summary of methods and features used by the teams.

7 Moldavian vs. Romanian Cross-dialect Topic identification (MRC)

Romanian (RO) is the language currently spoken in Romania, which is part of the Balkan-Romance group of languages. Besides Romanian, the group contains three other dialects: Aromanian, Istro-Romanian, and Megleno-Romanian. In order to distinguish Romanian within the Balkan-Romance group in comparative linguistics, it is referred to as *Daco-Romanian*. Moldavian (MD) is a subdialect of Daco-Romanian, that is spoken in the Republic of Moldova and in northeastern Romania. The delimitation of the Moldavian (sub)dialect, as with all other Romanian (sub)dialects, is mainly based on phonetic features and only marginally by morphological, syntactical, and lexical characteristics. Although the spoken dialects in Romania and the Republic of Moldova are different, the two countries share the same literary standard (Minahan, 2013). Some linguists (Pavel, 2008) consider that the border between Romania and the Republic of Moldova does not correspond to any significant isoglosses to justify a dialectal division. Therefore, separating between Romanian and Moldavian is a challenging task. The aim of the MRC shared task is (i) to determine to what the extent the two (sub)dialects can be automatically distinguished and (ii) to assess the performance of applying machine learning models trained on one dialect, e.g. Moldavian, directly (without fine-tuning) to the other, e.g. Romanian.

Set	#samples	#tokens
Training	21,719	6,705,334
Development	11,845	3,677,795
Private Test	5,923	1,836,705
Total	**39,487**	**12,219,834**

Table 5: The number of samples (#samples) and the number of tokens (#tokens) contained in the training, development (public validation plus test sets) and private test sets included in the MOROCO dataset.

7.1 Dataset

The dataset used in the MRC shared task was recently introduced in (Butnaru and Ionescu, 2019). The publicly available corpus[4], released before the MRC shared task, contains 33,564 samples collected from the news domains in Romania and Re-

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

public of Moldova. The samples belong to one of the following six topics: culture, finance, politics, science, sports, and tech. For the competition, we provided a distinct and private test set of 5,923 samples. The public validation and test sets we unified into a single development set for the competition. Table 5 provides the number of samples and the number of tokens in each subset (training, development and private test). The whole corpus is formed of 39,487 samples with over 12 million tokens. Since we provide both dialectal and category labels for each sample, we proposed three subtasks for the competition:

- Binary classification by dialect (subtask 1) – the task is to discriminate between the Moldavian and the Romanian dialects.
- MD→RO cross-dialect multi-class categorization by topic (subtask 2) – the task is to classify the samples written in the Romanian dialect into six topics, using a model trained on samples written in the Moldavian dialect.
- RO→MD cross-dialect multi-class categorization by topic (subtask 3) – the task is to classify the samples written in the Moldavian dialect into six topics, using a model trained on samples written in the Romanian dialect.

7.2 Participants and Approaches

DTeam. The approach of DTeam is based on an ensemble model that combines two character-level convolutional neural networks (CNN) using Support Vector Machines (SVM). The first CNN is based on a skip-gram model that is trained using softmax loss. The second CNN is trained using triplet loss (Schroff et al., 2015). DTeam submitted a single run to each of the three MRC subtasks.

lonewolf. The lonewolf team submitted three runs for subtask 1. The first run is based on a character-level bigram classification model to discriminate between Moldavian and Romanian examples using Add-One Smoothing for out-of-vocabulary items. The second and the third runs are based on word-level bigram classification models. The second run uses Add-One Smoothing for out-of-vocabulary items, while the third run uses Good-Turing Smoothing.

R2I_LIS. The R2I_LIS team submitted three runs for subtask 1. All their runs are based on a set of 40 features that include: the average length of a token, the average number of tokens per sentence, the number of tokens inside each text document, the number of occurrences of selected single characters, the number of occurrences of selected punctuation characters, the number of occurrences of the letter 'î' inside a word (not as the first character), the number of occurrences of selected words and the number of occurrences of the token NE which replaces named entities. The third run uses a normalized version of these features. All runs are based on a majority voting scheme applied on five classification models: k-Nearest Neighbors, Logistic Regression, Support Vector Machines, Neural Networks and Random Forests. For the first and the third runs, the models are trained on both training and development sets. For the second run, the model is trained only on the training set.

SC-UPB. The SC-UPB team first cleaned the dataset by removing stopwords as well as special characters. The first run submitted to each of the three subtasks is based on a model that represents text as the mean of word vectors given by a pre-trained FastText model (Grave et al., 2018). The representation is provided as input to a Recurrent Neural Network with gated recurrent units, which is trained using the Adam optimizer with a batch size of 64 for 20 epochs and early stopping. The second run submitted to each of the three subtasks is based on a hierarchical attention network introduced by Yang et al. (2016). The model is trained using the Adam optimizer with a batch size of 64 for 20 epochs and early stopping.

tearsofjoy. The tearsofjoy team used a linear SVM classifier with a combination of character and word *n*-gram features, which are weighted with the BM25 weighting scheme. Their model's parameters are tuned independently for each subtask, using random search and 5-fold cross-validation. The tearsofjoy team also tried a transductive learning approach which is based on re-training the model by adding confident predictions from the test set to the training set, an idea previously studied in (Ionescu and Butnaru, 2018).

7.3 Results

After the submission deadline, we noticed that two teams (tearsofjoy and SC-UPB) submitted runs containing less than the expected number of labels (5,923) for the test examples. Hence, their original (unmodified) submissions could not be eval-

Rank	Team	Run	F1 (macro)
1	DTeam	1	0.8950
2	R2I_LIS	3	0.7964
3	tearsofjoy	1	0.7573
4	lonewolf	2	0.7354
5	SC-UPB	1	0.7088

Table 6: Results on MRC subtask 1 (binary classification by dialect).

Rank	Team	Run	F1 (macro)
1	tearsofjoy	1	0.6115
2	SC-UPB	1	0.4813
3	DTeam	1	0.3856

Table 7: Results on MRC subtask 2 (multi-class categorization by topic of Romanian text samples using Moldavian text samples for training).

Rank	Team	Run	F1 (macro)
1	tearsofjoy	1	0.5533
2	SC-UPB	1	0.4808
3	DTeam	1	0.4472

Table 8: Results on MRC subtask 2 (multi-class categorization by topic of Moldavian text samples using Romanian text samples for training).

uated. In order to evaluate their runs, we tried to fix the problem by adding random labels using the following options: (i) at random locations in the submission files or (ii) at the end of the submission files. In the evaluation, we considered the option that provided better performance for the runs submitted by tearsofjoy and SC-UPB.

The best run of each participant in MRC subtask 1 is presented in Table 6. We notice that DTeam uses an approach based on deep learning, which surpasses the shallow approaches of R2I_LIS and tearsofjoy teams.

Table 7 contains the F1 (macro) score of the best run of each participant in MRC subtask 2. This time, we notice that the winning approach is shallow. It surpasses the other approaches based on deep neural networks. The ranking for subtask 2 is identical to the ranking for subtask 3, as shown in Table 8.

7.4 Summary

We proposed three MRC subtasks for VarDial 2019. Three participants submitted runs for all three subtasks, and another two participants submitted runs only for subtask 1. Two teams (DTeam and SC-UPB) proposed systems based on deep neural networks, while the other teams proposed shallow approaches based on handcrafted features. For subtask 1, the winning solution is a deep learning system. For subtasks 2 and 3, the winning solution is a shallow learning system. Hence, it remains unclear which of the two learning approaches, deep or shallow, provides better results in Moldavian vs. Romanian Cross-dialect Topic identification.

8 Cuneiform Language Identification (CLI)

The first edition of the CLI shared task was a language identification task concentrating on distinguishing between languages and dialects which were originally written with cuneiform signs. It included two completely separate languages: Akkadian and Sumerian. We had only one variety for Sumerian, but for Akkadian, we included six separate dialects: Old Babylonian, Middle Babylonian peripheral, Standard Babylonian, Neo-Babylonian, Late Babylonian, and Neo-Assyrian.

8.1 Dataset

The dataset used in the CLI shared task, as well as its creation, is described in detail by Jauhiainen et al. (2019a). The dataset was created using openly available transliterations originating from the Open Richly Annotated Cuneiform Corpus (Oracc). [5] In Oracc, the texts, originally written using the cuneiform script, are mostly stored in transliterated form. A special conversion program was used to transform these transliterated texts to Unicode cuneiform encoding. The data consists of texts originally appearing in one line of cuneiform writing. Word boundaries were not marked in the original script, but in the transliterations the word boundaries were marked. In order to produce more realistic cuneiform writing, the word boundaries were again removed in the conversion to Unicode cuneiform. Each line, thus, may consist of one or more words.

The sizes of the training sets for each language varied, and the exact number of lines in each can be seen in Table 9. In addition to the training set, the participants were provided with 668 lines of development data for each language. The test set had 985 lines for each language.

[5] http://oracc.museum.upenn.edu

Language or Dialect	Training
Sumerian	53,673
Old Babylonian	3,803
Middle Babylonian peripheral	5,508
Standard Babylonian	17,817
Neo-Babylonian	9,707
Late Babylonian	15,947
Neo-Assyrian	32,966

Table 9: Number of lines for each language or dialect in the CLI training set.

8.2 Participants and Approaches

In addition to the best performing system from each team, we have collected information about some of their other submissions if the systems used were clearly different. This information can be seen together with the test results in Table 10. The baseline methods and their results included in the table are described by Jauhiainen et al. (2019a).

The **NRC-CNRC** team submitted three runs. Their first submission was based on SVMs using character *n*-grams with different weighting schemes. Their second submission used a voting ensemble comprised of the previous SVMs and probabilistic classifiers. Their third and winning submission was based on a deep neural network (modified version of the BERT model) taking characters as input. With the deep neural network they had a second pre-training phase in which an unsupervised method was used to learn information from, and in a way adapt, to the test set. For more detailed information see the description by Bernier-Colborne et al. (2019).

The **tearsofjoy** team submitted two runs using SVMs. The better of their runs had two stages. After the first stage, those lines claimed by only one of the one-vs-all classifiers were added to the training data. This functions as one iteration of language model (LM) adaptation similar to the one used by Jauhiainen et al. (2018b) in the 2018 Indo-Aryan language identification (ILI) shared task. However, using language model adaptation improved their F1-score only by 1.6%. Their system is better described by Wu et al. (2019).

The **TwistBytes** team submitted two runs using SVMs. The better of their runs used tf-idf features with binary tf values and smoothed idf for character *n*-grams 1–3. Benites et al. (2019) describe the two systems in more detail.

The **PZ** team used a SVM metaclassifier ensemble of several linear SVM classifiers trained using character *n*-gram and character skip-gram features. Paetzold and Zampieri (2019) give further details.

The **SharifCL** team submitted three runs and their best performing system was an ensemble of a SVM and a NB classifier (Doostmohammadi and Nassajian, 2019).

The **ghpaetzold** team submitted only one run using 2-layer compositional recurrent neural network that learns numerical representations of sentences based on their words, and of words based on their characters. Their system is described in more detail by Paetzold and Zampieri (2019).

The **ekh** team used a sum of relative frequencies of character bigrams together with a penalty value for those bigrams or unigrams that were not found from a language.

The **situx** team used a Random Forest classifier. Their results are below a random baseline and we suspect there might have been some processing problems when generating the results from test set.

8.3 Results

Table 10 shows the performance of different methods on the CLI dataset. The ranked results are bolded in the table. To the best of our knowledge, this is the first time a language identification shared task has been won using neural networks in addition to the first MRC subtask.

8.4 Summary

We were happy to see such a widespread interest in the CLI shared task. The NRC-CNRC team did not participate in the other shared tasks, so we cannot directly compare the performance of their deep neural network between different writing systems. The only other team using neural networks was the ghpaetzold and the performance of their RNN is more in line what we have used to expect from neural networks when compared with the SVMs.

The second ranking team, tearsofjoy, used LM adaptation on the test set. They did the same with the GDI and the DMT tasks and were ranked very high in them as well. The difference in F1 score between their adaptive and non-adaptive systems is surprisingly small in CLI, as the test data in CLI was supposed to be out-of-domain when compared with the training and the development sets (Jauhiainen et al., 2019a).

Rank	Team	Method	Features used	F1
1.	**NRC-CNRC**	**Deep Neural Network + adapt.**	**characters**	**0.7695**
2.	**tearsofjoy**	**Lin. SVM with LM adapt.**	**ch. *n*-grams 1–5**	**0.7632**
	tearsofjoy	Lin. SVM	ch. *n*-grams 1–4	0.7511
	NRC-CNRC	SVM + NB ensemble	ch. *n*-grams 1–5	0.7449
3.	**Twist Bytes**	**Lin. SVM**	**ch. *n*-grams 1–3**	**0.7433**
	NRC-CNRC	SVM	ch. *n*-grams 1–4	0.7414
4.	**PZ**	**SVM ensemble**	**ch. *n*-grams 1–5, skip-grams**	**0.7347**
5.	**SharifCL**	**SVM + NB ensemble**	**ch. *n*-grams 1–4**	**0.7210**
	Baseline-3	Prod. of rel. freq.	ch. *n*-grams 1–4	0.7206
	SharifCL	SVM	ch. *n*-grams 1–4	0.7171
	Baseline-4	Voting ensemble	ch. *n*-grams 1–15	0.7163
	Baseline-5	HeLI	ch. *n*-grams 1–3 + lines	0.7061
	Twist Bytes	Lin. SVM	ch. *n*-grams 1–7, words	0.6669
	Baseline-1	Simple scoring	ch. *n*-grams 1–10	0.6554
	Baseline-2	Sum of rel. freq.	ch. *n*-grams 3–15	0.6016
6.	**ghpaetzold**	**RNN**	**characters, words**	**0.5562**
7.	**ekh**	**Sum of rel. freq. + spec. penalt.**	**ch. 2-grams**	**0.5501**
8.	**situx**	**Random Forest**	**ch. *n*-grams 2–4, spec.**	**0.1276**

Table 10: The macro F1-scores attained by the participating teams and the baseline methods with the CLI dataset. The official ranked results are bolded.

9 Conclusion

In this paper, we presented the results and the main findings for the five shared tasks organized as part of the Third VarDial Evaluation Campaign. One task was a re-run from previous years (GDI), and four new tasks were organized: CMA, DMT, MRC, and CLI.

A total of 22 teams submitted runs across the five shared tasks. We included short descriptions for each participant's systems in this report. A complete description is available in the system description papers, which were presented in the VarDial workshop and published in the workshop proceedings. We included references to all system description papers in this report in Table 1.

Acknowledgments

We would like to thank the participants of the Third VarDial Evaluation Campaign for their participation, support, and feedback. We further thank the VarDial workshop program committee members for thoroughly reviewing the shared task system description papers and this report.

The GDI organizers would like to thank Thayabaran Kathiresan and Lei He (University of Zurich) for their valuable help with the preparation of the iVectors.

The CLI organizer is grateful to his colleagues at the Centre of Excellence in Ancient Near Eastern Empires (ANEE) of the University of Helsinki for their invaluable assistance.

References

Ahmed Ali, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell, and Steve Renals. 2016. Automatic Dialect Detection in Arabic Broadcast Speech. In *Proceedings of INTERSPEECH*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Fernando Benites, Pius von Däniken, and Mark Cieliebak. 2019. TwistBytes - Identification of Cuneinform Languages and German Dialects at VarDial 2019. In *Proceedings of VarDial*.

Gabriel Bernier-Colborne, Cyril Goutte, and Serge Léger. 2019. Improving Cuneiform Language Identification with BERT. In *Proceedings of VarDial*.

Andrei Butnaru and Radu Tudor Ionescu. 2019. MOROCO: The Moldavian and Romanian Dialectal Corpus. *arXiv preprint arXiv:1901.06543*.

Andrei M. Butnaru. 2019. BAM: A combination of deep and shallow models for German Dialect Identification. In *Proceedings of VarDial*.

Çağrı Çöltekin and Jeremy Barnes. 2019. Neural and Linear Pipeline Approaches to Cross-lingual Morphological Analysis. In *Proceedings of VarDial*.

Keh-Jiann Chen, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. 1996. Sinica Corpus : Design Methodology for Balanced Corpora. In *Proceedings of PACLIC*.

Adrian-Gabriel Chifu. 2019. The R2I_LIS Team Proposes Majority Vote for VarDial's MRC Task. In *Proceedings of VarDial*.

Najim Dehak, Pedro A. Torres-Carrasquillo, Douglas Reynolds, and Reda Dehak. 2011. Language recognition via i-vectors and dimensionality reduction. In *Proceedings of INTERSPEECH*.

Eugen Dieth. 1986. *Schwyzertütschi Dialäktschrift*, 2 edition. Sauerländer.

Ehsan Doostmohammadi and Minoo Nassajian. 2019. Investigating Machine Learning Methods for Language and Dialect Identification of Cuneiform Texts . In *Proceedings of VarDial*.

Mikel Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25:127–144.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of LREC*.

Renkui Hou and Chu-Ren Huang. 2019. (to appear) Classification of Regional and Genre Varieties of Chinese: A correspondence analysis approach based on comparable balanced corpora. *Journal of Natural Language Engineering*.

Hai Hu, Wen Li, He Zhou, Zuoyu Tian, Yiwen Zhang, and Liang Zou. 2019. Ensemble Methods to Distinguish Mainland and Taiwan Chinese. In *Proceedings of VarDial*.

Chu-Ren Huang and Dingxu Shi. 2016. *A Reference Grammar of Chinese*. Cambridge University Press.

Radu Tudor Ionescu and Andrei M. Butnaru. 2018. Improving the results of string kernels in sentiment analysis and Arabic dialect identification by adapting them to your test set. In *Proceedings of EMNLP*.

Tommi Jauhiainen, Heidi Jauhiainen, Tero Alstola, and Krister Lindén. 2019a. Language and Dialect Identification of Cuneiform Texts. In *Proceedings of VarDial*.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2018a. HeLI-based experiments in Swiss German dialect identification. In *Proceedings of VarDial*.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2018b. Iterative language model adaptation for Indo-Aryan language identification. In *Proceedings of VarDial*.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2019b. Discriminating between Mandarin Chinese and Swiss-German varieties using adaptive language models . In *Proceedings of VarDial*.

Jingxia Lin, Dingxu Shi, Menghan Jiang, and Chu-Ren Huang. 2018. Variations in World Chineses. *The Routledge Handbook of Applied Chinese Linguistics*.

Krister Lindén. 2009. Guessers for finite-state transducer lexicons. In *Proceedings of CICLing*.

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*.

A. M. McEnery and R. Z. Xiao. 2003. The Lancaster Corpus of Mandarin Chinese.

Vladislav Mikhailov, Lorenzo Tosi, Anastasia Khorosheva, and Oleg Serikov. 2019. Initial Experiments In Cross-Lingual Morphological Analysis Using Morpheme Segmentation. In *Proceedings of VarDial*.

James Minahan. 2013. *Miniature Empires: A Historical Dictionary of the Newly Independent States*. Routledge.

Sarah Moeller, Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2018. A neural morphological analyzer for Arapaho verbs learned from a finite state transducer. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*.

Garrett Nicolai and Grzegorz Kondrak. 2017. Morphological analysis without expert annotation. In *Proceedings of EACL*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Chris Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Dan Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of LREC*.

Cristian Onose and Dumitru-Clementin Cercel. 2019. SC-UPB Team at the VarDial 2019 Evaluation Campaign: Moldavian vs. Romanian Cross-Dialect Topic Identification. In *Proceedings of VarDial*.

Gustavo H. Paetzold and Marcos Zampieri. 2019. Experiments in Cuneiform Language Identification. In *Proceedings of VarDial*.

Vasile Pavel. 2008. Limba română – unitate în diversitate (Romanian language – there is unity in diversity). *Romanian Laguage Journal*, XVIII(9–10).

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

Yves Scherrer and Nikola Ljubešić. 2016. Automatic Normalisation of the Swiss German ArchiMob Corpus Using Character-level Machine Translation. In *Proceedings of KONVENS*.

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. in press. Digitising Swiss German – how to process and study a polycentric spoken language. *Language Resources and Evaluation*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of CVPR*.

Miikka Silfverberg and Mans Hulden. 2018. Initial experiments in data-driven morphological analysis for Finnish. In *Proceedings of the Fourth International Workshop on Computatinal Linguistics of Uralic Languages*.

Miikka Silfverberg and Francis Tyers. 2019. Data-Driven Morphological Analysis for Uralic Languages. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, Mikko Kurimo, et al. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of EACL*.

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*.

Diana-Elena Tudoreanu. 2019. DTeam @ VarDial 2019: Ensemble based on skip-gram and triplet loss neural networks for Moldavian vs. Romanian cross-dialect topic identification. In *Proceedings of VarDial*.

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*.

Li Yang and Yang Xiang. 2019. Naive Bayes and BiLSTM Ensemble for Discriminating between Mainland and Taiwan Variation of Mandarin Chinese. In *Proceedings of VarDial*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of NAACL*.

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

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