Applying Multilingual and Monolingual Transformer-Based Models for Dialect Identification

Cristian Popa, Vlad Ștefănescu
University Politehnica of Bucharest
{cristian.viorel.popa, vlad.a.stefanescu}@gmail.com

Abstract

We study the ability of large fine-tuned transformer models to solve a binary classification task of dialect identification, with a special interest in comparing the performance of multilingual to monolingual ones. The corpus analyzed contains Romanian and Moldavian samples from the news domain, as well as tweets for assessing the performance. We find that the monolingual models are superior to the multilingual ones and the best results are obtained using an SVM ensemble of 5 different transformer-based models. We provide our experimental results and an analysis of the attention mechanisms of the best-performing individual classifiers to explain their decisions. The code we used was released under an open-source license.

1 Introduction

Dialect Identification is a Natural Language Processing (NLP) task that started receiving more interest in recent years, in part due to VarDial, the workshop on NLP for Similar Languages, Varieties and Dialects (Nakov et al., 2017; Zampieri et al., 2018b; Zampieri et al., 2019b) and its organized evaluation campaigns (Zampieri et al., 2017; Zampieri et al., 2018a; Zampieri et al., 2019a).

This paper presents our solution to the RDI shared task of VarDial 2020 (Găman et al., 2020) on behalf of team "Anumiți". The problem we focus on is the binary classification of Moldavian and Romanian samples, training on the MOROCO (Butnaru and Ionescu, 2019) corpus and a small number of collected tweets, and testing afterwards on an additional set of tweets. Multiple experiments were done on this corpus (Găman and Ionescu, 2020; Tudoreanu, 2019; Chifu, 2019; Wu et al., 2019; Onose et al., 2019), but we are, to the best of our knowledge, the first ones to apply some of the models in this paper to solve the task.

We study the performance of 3 multilingual models, trained on Romanian corpora, and 2 other BERT-based models trained only on Romanian data. The results show clearly that the ones explicitly trained for Romanian tasks are superior for the dialect identification task at hand. We present the results of the different individual classifiers and an SVM ensemble of them, as well as the words that the Romanian-trained ones pay attention to, in order to differentiate between the two dialects. We released all the code under an open-source license: https://github.com/CristianViorelPopa/transformers-dialect-identification.

2 Related Work

Considerable progress was made in the NLP field in recent years with the introduction of attention models (Bahdanau et al., 2014) and, later on, transformers (Vaswani et al., 2017), deep neural networks that use an encoder-decoder architecture. A multitude of models (Devlin et al., 2019; Radford et al., 2019; Liu et al., 2019; Lewis et al., 2019) that make use of the transformer architecture emerged and achieved state-of-the-art performance on a large number of NLP tasks through transfer learning. These are pre-trained on large amounts of textual data in order to encompass general knowledge of the target language(s) to later be fine-tuned on downstream tasks, such as question-answering (Rajpurkar et al.,

Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 193–201
Barcelona, Spain (Online), December 13, 2020

2016). This allows for better results on small datasets, where previously deep learning was not a viable solution. The architecture of the transformer-based models features multiple layers of transformers with multi-head attention, which allow the model to jointly attend to information from different representation subspaces at different positions (Vaswani et al., 2017).

A short time after, multilingual transformer-based models made an appearance. These are trained on multiple languages in order to solve suitable problems, such as cross-lingual sentence classification (Conneau et al., 2018). The multilingual BERT model (*mBERT*) uses the original BERT architecture and training objectives, but is trained on corpora of up to 104 languages, including Romanian. The model learns to infer masked tokens in the input sequences (*Masked Language Modeling*) and ignores the binary classification (*Next Sentence Prediction*) objective of the initial BERT. *XLM* (Lample and Conneau, 2019) is a cross-lingual model employs multiple language modeling objectives, including a modified variant of the BERT Masked Language Modeling supporting text streams with an arbitrary number of sentences and subsampling of frequent tokens. *XLM-RoBERTa* (*XLM-R*) (Conneau et al., 2019) is an update of XLM that is additionally trained on 2.5TB of newly cleaned CommonCrawl (Wenzek et al., 2019) data. Both XLM and XLM-R are trained on 100 languages, including Romanian.

In addition to the multilingual models, there is also the possibility of training transformer models, such as BERT, from scratch using only monolingual data for the specific task. Such progress has been made in the case of the Romanian language (Dumitrescu et al., 2020). These are trained on the OPUS (Tiedemann and Nygaard, 2004), OSCAR (Suárez et al., 2019) and Romanian Wikipedia corpora.

The popularity of transformer-based models determined a lot of research to be done for explaining their inside attention mechanisms (Tenney et al., 2019; Michel et al., 2019; Rogers et al., 2020; Clark et al., 2019). This growing field of study is called "BERTology."

3 Method

Our research focuses entirely on pre-trained transformer-based models fine-tuned using the same textual data and similar hyper-parameters.

3.1 Corpus Pre-processing

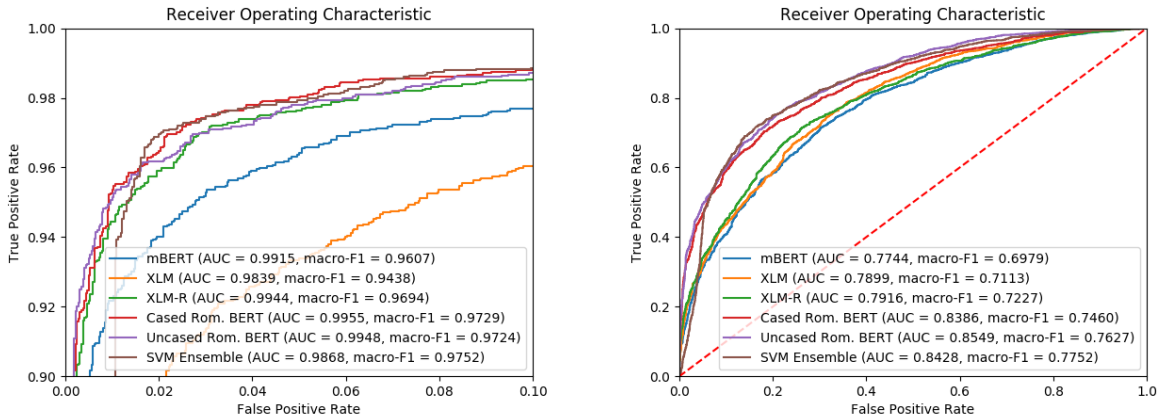
The training, validation and test datasets have two distinct categories of data: extracts from the news domain and tweets. The training set contains 33,564 news extracts and no tweets, the validation set has 5,923 news extracts and 215 tweets, while the test set consists of 5,022 tweets and no news extracts.

We apply multiple pre-processing techniques on these datasets. First of all, we remove recurring JavaScript artifacts. This only applies for the news samples. Further on, we attempt to remove unnecessary whitespaces, such as the ones before commas and periods. Additionally, we normalize the usage of punctuation, as there are multiple similar ASCII characters for the same punctuation symbol and there is the possibility of one symbol being equivalent to multiple ones (e.g. "…" as a single ASCII symbol and "..." as three period symbols).

The named entities in the given corpus are all replaced by the tokens "NE." We make the decision to replace these with "[MASK]" tokens, which are relevant in the language modeling objective of models such as BERT.

Finally, the two types of data are very distinct in regards to their sample size, with news extracts in the training set having an average character length of 1715, while the tweets in the test set have an average length of 92. For this reason, we observed that using the training data in its original state does not achieve satisfying results on tweets. To fix this, we split each news extract into sentences.[1] This brings the number of training samples to 366,628, more than 10 times the initial size of the dataset. We showcase the huge boost in performance this attains in a later section.

[1]To do this, we use the sentence segmentation available in the spaCy package (Honnibal and Montani, 2017)



(a) Performance of all models on news extracts.

(b) Performance of all models on tweets.

Figure 1: Performance comparison between all transformer models and the SVM ensemble.

3.2 Training the Models

We use 3 multilingual models and 2 monolingual models in our experiments. The multilingual models are the base (12-layer, 12-heads) cased variant of mBERT[2], the 16-layer, 16-heads XLM trained on 100 languages, and the large (24-layer, 16-heads) XLM-R[3]. The monolingual models use the BERT architecture, are trained only on Romanian data and employ 12 layers and 12 heads, similar to the mBERT model (Dumitrescu et al., 2020). They vary by the data tokenizer they use (cased vs. uncased).[4] We refer to the latter as cased/uncased Romanian BERT.

There are potential advantages and disadvantages to using both monolingual and multilingual models. The former are more specifically trained for the task, while the latter may encompass better linguistic knowledge. Nonetheless, neither the multilingual models, nor the monolingual ones do not explicitly incorporate Moldavian corpora during their initial training.

For the dialect identification task, an additional classification layer is added. All the models are fine-tuned under similar conditions. We use an Adam optimizer (Kingma and Ba, 2014) with a learning rate of 2e-5 and a warmup of 6% of total steps for 2 epochs. The input samples are truncated to 128 tokens. Note that this favors our initial decision of splitting the news extracts into sentences, as a large chunk of information would be lost otherwise. Due to hardware limitations, the XLM and XLM-R models use a batch size of 16, compared to the 32 of the other models, as these are significantly larger.[5] This is the only difference in hyper-parameters for the fine-tuning. The models are only trained on the news extracts in the training set. We choose the decision threshold for each model to be the one that maximizes the macro-F1 score.

A final SVM ensemble is trained from the probabilities generated by the 5 transformer-based models on every sample. We split the 215 tweets in the validation set into 108 samples that we use to train the ensemble and 107 samples to test its performance and determine the optimal threshold with the same objective as before. The best hyper-parameters of the SVM are found using a simple grid search, given the low dimensionality of the dataset, comprising only 5 numerical features. The kernel chosen following it was the radial basis function.

4 Experimental Results

In the results presented further, the metrics we analyze are the Area Under Curve (AUC) and macro-averaged F1 score, which requires a decision threshold to clearly distinguish Romanian-predicted sam-

[2] https://github.com/google-research/bert/blob/master/multilingual.md
[3] https://github.com/facebookresearch/XLM
[4] https://github.com/dumitrescustefan/Romanian-Transformers
[5] We use the HuggingFace Transformers (Wolf et al., 2019) library for our experiments.

Table 1: Qualitative results of all the models used.

Models	**News Extracts**		**Tweets**	
	AUC	Macro-F1	AUC	Macro-F1
mBERT	0.9915	0.9607	0.7744	0.6979
XLM	0.9839	0.9438	0.7899	0.7113
XLM-R	0.9944	0.9694	0.7916	0.7227
Cased Rom. BERT	**0.9955**	0.9729	0.8386	0.7460
Uncased Rom. BERT	0.9948	0.9724	**0.8549**	0.7627
SVM Ensemble	0.9868	**0.9752**	0.8428	**0.7752**

Table 2: Best macro-F1 scores for all the submissions on the RDI shared task.

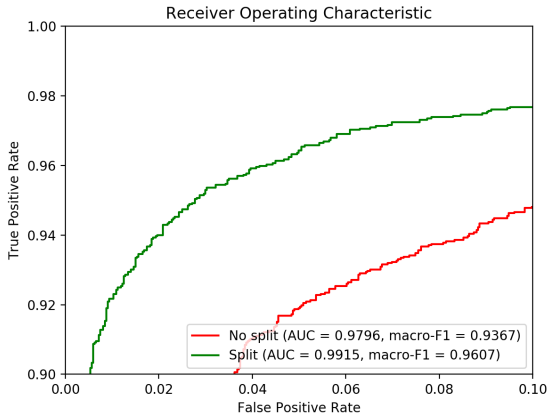
Models	**Macro-F1**
Tubingen	0.787592
Anumiți (SVM Ensemble)	**0.775178**
Phlyers	0.666090
SUKI	0.658437
UPB	0.647577
UAIC	0.555044
akanksha	0.481325
The Linguistadors	0.429412

ples from Moldavian ones. We choose to study the AUC since we believe it is stronger than the F1 score for showcasing the underlying potential of the model, as it is based on all the possible decision thresholds. The news extracts and tweets have distinct enough samples that the performance varies greatly. For this reason, we consider both results relevant and showcase them separately. For the news extracts, we assess the performance on the 5,923 extracts in the validation set, with the threshold being set based on the same samples. For the tweets, we determine the threshold based on the 215 samples in the validation set, while we are interested in the performance of the larger test set with 5,022 samples. The SVM ensemble is an exception to this rule, as it is trained on half of the validation tweets, while the other half is used to determine the best threshold.

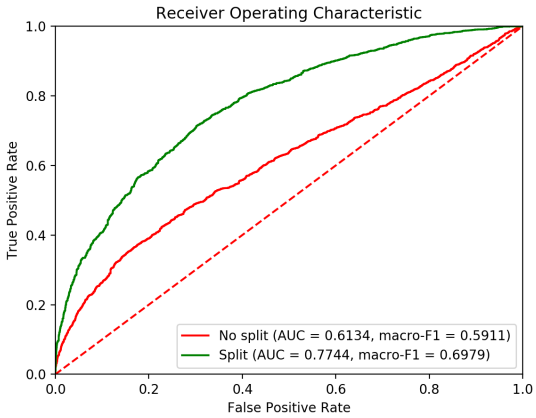
Because the performance on the news extracts is very good and the AUC quickly approaches 1.0 for all models, we only showcase a specific area of the ROC with False Positive Rate (FPR) between 0.0 and 0.1, as well as True Positive Rate (TPR) between 0.9 and 1.0. This way, we are able to present a clearer comparison between the models.

4.1 Comparison of the Models

First of all, we investigate the performance of all the individual models and the SVM ensemble on the two types of datasets. The ROC of each one can be seen in Figure 1 and the numeric values of the metrics are recorded in Table 1. We notice some clear differences in behaviour between the two cases. For the news extracts, the two Romanian variants of BERT and XLM-R are closely competing, while mBERT is performing slightly worse and XLM is the obvious loser. In regards to the tweets, we see that the monolingual models are superior to all the multilingual ones, with the uncased version coming out on top. In both cases, the SVM ensemble is not the best according to the AUC, but still achieves the



(a) Performance of mBERT on news extracts.



(b) Performance of mBERT on tweets.

Figure 2: Performance comparison for mBERT trained on split vs. unsplit samples.

highest macro-F1 score. The three runs we submitted were the ones with the highest macro-F1 scores on the "Tweets" set, namely the two Romanian BERT models, along with the SVM ensemble. Table 2 displays the best macro-F1 scores obtained by the teams participating in the RDI shared task of VarDial 2020. We closely trail the first place submission, validating that transformer-based models are viable and competitive in classifying dialects.

Given that the Romanian BERT models achieve the best results on both the news extracts and the tweets, we may conclude that for identifying between Romanian and Moldavian dialects, monolingual models trained on Romanian corpora are better than their multilingual counterparts trained on a large number of languages.

4.2 Improvement of Splitting

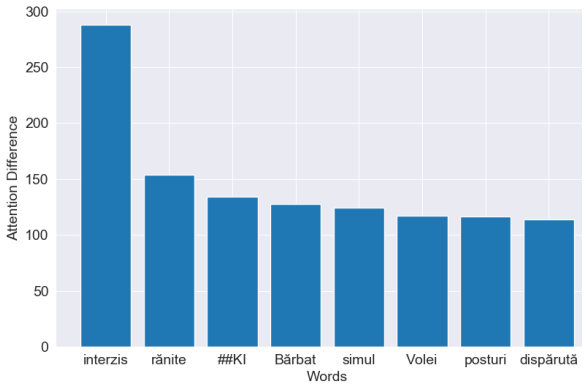
As mentioned previously, the decision to split the extracts from the news domain into sentences and use those instead as training samples brought a huge improvement to our models. We showcase this with the mBERT model in Figure 2. Notice that the performance on the tweets was greatly enhanced and there is an additional small boost on the news extracts as well.

4.3 Analysis of the Attention Mechanism

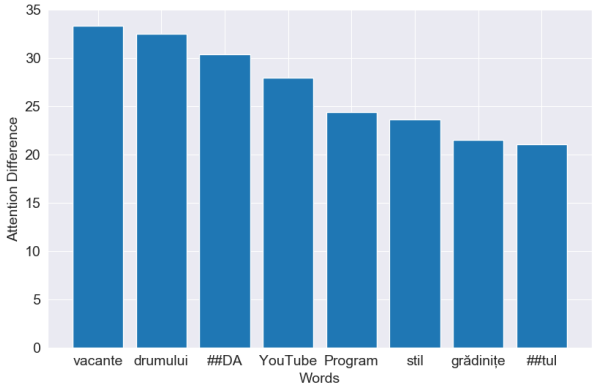
Inspired by the works of Clark et al. (2019), we attempt to explain the decisions of some of our models, namely the monolingual ones, as they proved to achieve the best results. We are interested in finding the tokens most important in differentiating between the two dialects, based on the attention that is being paid to them inside the transformer modules. The encoder in the transformer architecture (Vaswani et al., 2017) features multiple attention heads where each token in the input attends to the others with a certain degree of attention. Each layer in the BERT architecture is such an encoder employing multi-head attention. In our case, the monolingual models comprise 12 layers of encoders, each one having 12 attention heads.

We define the importance of a token in classifying a sample as the total amount of attention that is being paid to it over all layers and attention heads, divided by the number of instances it is being encountered in. Mathematically, the importance of a token t_{target} in a corpus C it is defined as follows:

$$Importance(t_{target}) = \frac{\sum_{i=1}^{n_l}\sum_{j=1}^{n_h}\sum_{s\in C}\sum_{t_a\in s}\sum_{t_b\in s}\begin{cases} attn_{ij}(t_a,t_b), t_b = t_{target} \\ 0, t_b \neq t_{target} \end{cases}}{\sum_{i=1}^{n_l}\sum_{j=1}^{n_h}\sum_{s\in C}\sum_{t_a\in s}\sum_{t_b\in s}\begin{cases} 1, t_b = t_{target} \\ 0, t_b \neq t_{target} \end{cases}} \quad (1)$$

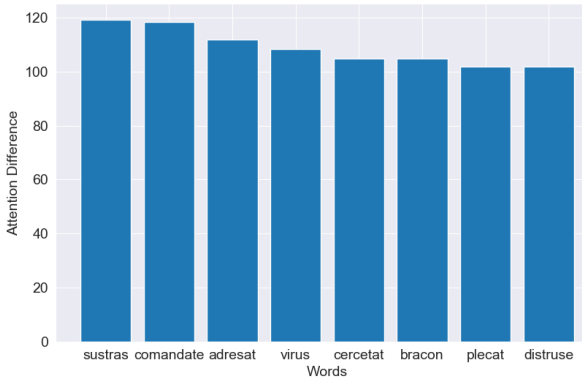


(a) Differentiating tokens for the Romanian dialect.

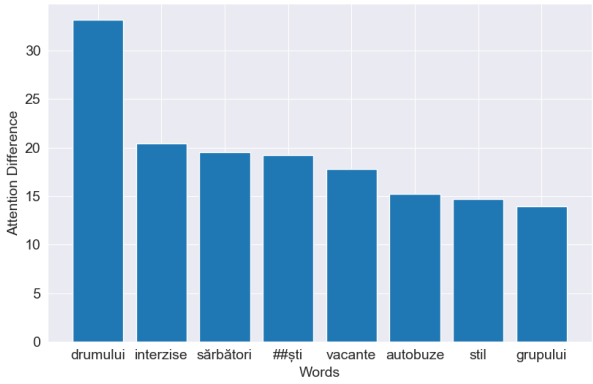


(b) Differentiating tokens for the Moldavian dialect.

Figure 3: Differentiating tokens for both dialects extracted from the cased Romanian BERT.



(a) Differentiating tokens for the Romanian dialect.



(b) Differentiating tokens for the Moldavian dialect.

Figure 4: Differentiating tokens for both dialects extracted from the uncased Romanian BERT.

where $attn_{ij}(t_a, t_b)$ is the attention that token t_a pays to token t_b in layer i and head j, n_l is the total number of layers and n_h is the number of attention heads per layer.

The importance of a token is computed for Romanian and Moldavian samples separately. The difference between them for a single token represents its capacity to distinguish the Romanian dialect from the Moldavian one (Romanian-biased) and vice-versa (Moldavian-biased). We refer to these tokens as "*differentiating tokens*" and extract them only from the tweets in the test set. Instead of the real labels, we choose to use the ones predicted by the models. The plots in Figures 3 and 4 showcase the top-8 differentiating tokens found in all cases and their corresponding difference in attention between the two dialects. The "##" symbols indicate that the token is a continuation of a previous one for the same word (e.g. "*##face*" for "*preface*"). Table 3 contains the number of appearances for each of these tokens in the Romanian and Moldavian samples.

One first observation that can be made is that the attention being paid to Romanian-biased differentiating tokens has a higher magnitude than the Moldavian-biased ones. As we expect, with few exceptions, the tokens are encountered multiple times in the dialect they are biased towards and usually once in the other dialect (note that we only consider tokens present in both). This means that the models are able to pay attention to meaningful tokens, that are favored by one dialect, even though these may well be used by both, such as "*virus*." As native speakers of the Romanian language, with minor knowledge of Moldavian-specific constructs, we don't find these tokens particularly related to grammar, but rather to trends in the media. For instance, the token "*Volei*" may be prevalent in Romanian samples due to the

Table 3: Differentiating tokens for the Romanian BERT models (top half - tokens biased toward Romanian samples; bottom half - tokens biased toward Moldavian samples; left half - cased version; right half - uncased version) and their corresponding number of appearances in the corpora, based on predicted label.

Tokens	**RO Count**	**MD Count**	**Tokens**	**RO Count**	**MD Count**
interzis	1	3	sustras	9	2
rănite	9	2	comandate	1	1
##KI	3	1	adresat	1	1
Bărbat	12	1	virus	5	1
simul	1	1	cercetat	10	1
Volei	4	1	bracon	6	1
posturi	3	1	plecat	8	1
dispărută	2	1	distruse	1	1
vacante	1	5	drumului	1	5
drumului	1	5	interzise	1	4
##DA	2	36	sărbători	1	5
YouTube	6	96	##ști	1	3
Program	1	6	vacante	1	5
stil	1	7	autobuze	2	7
grădinițe	1	10	stil	1	7
##tul	1	4	grupului	1	12

popularity of their national Volleyball team. An interesting takeaway is that Romanian media tends to utilize more adjectives, compared to the Moldavian one making use of nouns for the most part. This indicates that expressiveness was favored in the Romanian samples.

5 Conclusion

In this paper, we analyzed a total of 3 multilingual and 2 monolingual transformer-based models fine-tuned on a task of dialect identification between Romanian and Moldavian dialects. We present our methodology for pre-processing the data and found that the monolingual models outclass the multilingual ones on data originating from two different sources, with an SVM ensemble combining them all that achieves the best macro-F1 score. Lastly, we provide a study on the explainability of the best models, focusing on the tokens that these classifiers find the most important in distinguishing the dialects.

References

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

Andrei Butnaru and Radu Tudor Ionescu. 2019. MOROCO: The Moldavian and Romanian dialectal corpus. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 688–698, Florence, Italy, July. Association for Computational Linguistics.

Adrian-Gabriel Chifu. 2019. The R2I_LIS team proposes majority vote for VarDial's MRC task. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 138–143, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look At? An Analysis of BERT's Attention. In *BlackBoxNLP@ACL*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv preprint arXiv:1911.02116*.

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, June. Association for Computational Linguistics.

Stefan Daniel Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT.

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

Mihaela Găman, 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 Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, pages 14014–14024.

Preslav Nakov, Marcos Zampieri, Nikola Ljubešić, Jörg Tiedemann, Shevin Malmasi, and Ahmed Ali, editors. 2017. *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Valencia, Spain, April. Association for Computational Linguistics.

Cristian Onose, Dumitru-Clementin Cercel, and Stefan Trausan-Matu. 2019. SC-UPB at the VarDial 2019 evaluation campaign: Moldavian vs. Romanian cross-dialect topic identification. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 172–177, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *arXiv preprint arXiv:2002.12327*.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July. Association for Computational Linguistics.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Diana 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 the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 202–208, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

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 the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 54–63, Ann Arbor, Michigan, June. 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, April. 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. 2018a. 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, August. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Nikola Ljubešić, Jörg Tiedemann, Shervin Malmasi, and Ahmed Ali, editors. 2018b. *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, Santa Fe, New Mexico, USA, August. 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. 2019a. 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, June. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Shervin Malmasi, Nikola Ljubešić, Jörg Tiedemann, and Ahmed Ali, editors. 2019b. *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, Ann Arbor, Michigan, June. Association for Computational Linguistics.