Social Media Variety Geolocation with geoBERT

Yves Scherrer
Department of Digital Humanities
University of Helsinki
yves.scherrer@helsinki.fi

Nikola Ljubešić
Department of Knowledge Technologies
Jožef Stefan Institute
nikola.ljubesic@ijs.si

Abstract

This paper describes the Helsinki–Ljubljana contribution to the VarDial 2021 shared task on social media variety geolocation. Following our successful participation at VarDial 2020, we again propose constrained and unconstrained systems based on the BERT architecture. In this paper, we report experiments with different tokenization settings and different pre-trained models, and we contrast our parameter-free regression approach with various classification schemes proposed by other participants at VarDial 2020. Both the code and the best-performing pre-trained models are made freely available.

1 Introduction

The SMG (Social Media Geolocation) task was first introduced at VarDial 2020 (Gaman et al., 2020). In this task, the participants were asked to predict latitude-longitude coordinate pairs. This contrasts with most other VarDial tasks, in which the goal is to choose from a finite set of variety labels. The second edition of the SMG task is run at VarDial 2021 (Chakravarthi et al., 2021), with the same three language areas as in the previous year: the Bosnian-Croatian-Montenegrin-Serbian (BCMS) language area, the German language area comprised of Germany and Austria (DE-AT), and German-speaking Switzerland (CH). All three datasets are based on social media data, Twitter in the case of BCMS (Ljubešić et al., 2016) and Jodel in the case of DE-AT and CH (Hovy and Purschke, 2018).

This paper describes the HeLju (Helsinki–Ljubljana) submission to the SMG task. Following our successful participation in 2020 (Scherrer and Ljubešić, 2020), we again propose systems based on the BERT architecture in both constrained and unconstrained settings. We report experiments with different tokenization parameters and with newly available pre-trained models. Furthermore, we compare our parameter-free double regression approach with three classification schemes proposed by other participants at VarDial 2020. The models and code used in our final submissions are available at https://github.com/clarinsi/geobert.

2 Related work

One of the first works focusing on predicting geolocation from social media text is Han et al. (2012). The authors investigate feature (token) selection methods for location prediction, showing that traditional predictive algorithms yield significantly better results if feature selection is performed.

There has been already a shared task on geolocation prediction at WNUT 2016 (Han et al., 2016). The task focused not only on predicting geolocation from text, but also from various user metadata. The best performing systems combined the available information via feedforward networks or ensembles.

Thomas and Hennig (2018) report significant improvements over the winner of the WNUT-16 shared task by separately learning text and metadata embeddings via different neural network architectures (LSTM, feedforward), merging those embeddings and performing the final classification via a softmax layer.

During the last iteration of the VarDial social media geolocation shared task (Gaman et al., 2020), most teams submitted constrained systems, i.e., they did not make use of any form of transfer learning from other tasks or pre-training self-supervision. Two approaches emerged during the shared task – a direct one, modelling the problem as multi-target regression, and an indirect one, which converts the coordinates into a finite set of dialect areas and uses a classification model to predict one of the areas. Interestingly, both approaches obtained similar results in the constrained setting, while only multi-target regression was tested in

Proceedings of the 8th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 135–140
April 20, 2021 ©2021 Association for Computational Linguistics

Instances	BCMS	CH	DE-AT
Training	353 953	25 261	318 487
Development	38 013	2 416	29 122
Test	4 189	2 438	31 515

Table 1: Data characteristics.

the unconstrained one. We will inspect last year's submissions in more detail in Section 4.2.

3 Data

The VarDial evaluation campaign provides training, development and test data for the three subtasks (Chakravarthi et al., 2021). Table 1 gives an overview of the data. It can be seen that the BCMS and DE-AT datasets are roughly equivalent in size, whereas the CH dataset is more than one order of magnitude smaller. The BCMS test set is much smaller than the development set, whereas the two are roughly equal in size in the Jodel-based subtasks. Note that the instances from the Twitter dataset (BCMS) correspond to single tweets, while the instances from the Jodel datasets correspond to entire conversations and are much longer on average.

The task organizers also provide a simple baseline for the geolocation task. They compute the centroid ("average location") of all instances in the training data and then predict the coordinates of this centroid for all development or test instances.

While the official metric of the shared task is median distance (in kilometers) between the gold and predicted geolocations, the organizers report back both the median and the mean distance metrics. In the previous iteration of the shared task, the two metrics showed to correlate in most, but not all cases.

4 Experiments

Due to lack of time (the two evaluation campaigns were held just a few months apart due to the *ACL conference bidding procedure), all our experiments are based on our successful 2020 submissions (Scherrer and Ljubešić, 2020): we use the BERT architecture (Devlin et al., 2019) with a fully connected layer on top of the CLS token. This fully connected layer implements double regression with a two-dimensional output vector and Mean Absolute Error loss.

We provide both constrained submissions, where BERT models are pre-trained from scratch using the VarDial training data, and unconstrained submissions, where we use pre-trained off-the-shelf BERT models. For all our experiments, we rely on the *simpletransformers* library, which is built on top of the HuggingFace *Transformers* library (Wolf et al., 2019), and use the same hyperparameters as in our 2020 submissions (Scherrer and Ljubešić, 2020).[1]

Task	Vocab. size	Median distance	Mean distance	Eps.
BCMS	3k	92.74	129.14	28
	30k	59.93	109.51	23
CH	3k	22.94	33.01	11
	30k	21.20	30.60	9
DE-AT	3k	182.76	205.90	4
	30k	160.67	186.35	2

Table 2: Effect of different vocabulary sizes on constrained model performance, evaluated on the development set. *Eps.* refers to the number of fine-tuning epochs to reach minimum median distance.

In the following subsections, we discuss some additional experiments carried out for the 2021 edition.

4.1 Tokenization

The BERT architecture requires a fixed vocabulary that is defined during the pre-training phase. Automatic word segmentation techniques are used to represent all tokens present in the data with a fixed-size vocabulary. Most off-the-shelf pretrained models are based on a vocabulary size of 30 000 units. However, we hypothesize that when dealing with language variation, a more aggressive split that approaches a character-by-character split might be beneficial.

On the basis of the constrained setup, we trained BERT models from scratch with vocabulary sizes of 3 000 and 30 000 units, with masked language modeling as pre-training task and geolocation pre-

[1] The maximum length of the instances was capped at 128 tokens for BCMS and CH and at 256 tokens for DE-AT. Intermediate models were saved every 2000 training steps and the savepoint with the lowest median distance value measured on the development set was selected for testing. For the regression models, we converted the coordinates using joint scaling and used the MAE loss. For the classification models, we used the default cross-entropy loss.

Pre-training from scratch is done with the masked language modelling task and default parameters.

Task	Approach and parameters	Number of classes	Train. reconstr. error (median)	Dev. reconstr. error (median)	Median dist.	Mean dist.	Eps.
BCMS	Regression	—	0	0	62.21	110.56	28
	K-means: $k = 35$	35	5.15	5.07	46.74	111.77	5
	K-means: $k = 75$	75	2.67	3.00	**41.84**	**107.38**	4
	Fixed cell numbers: 9×9	60	31.98	31.25	57.32	129.00	2
	Fixed cell size: $1° \times 1°$	38	39.92	36.68	46.92	119.79	3
CH	Regression	—	0	0	**21.20**	**30.60**	9
	K-means: $k = 35$	35	4.42	4.48	23.06	32.77	4
	K-means: $k = 75$	75	0.28	0.35	22.18	32.48	3
	Fixed cell numbers: 9×9	55	8.36	8.36	21.69	33.56	3
	Fixed cell size: $0.2° \times 0.2°$	61	7.52	7.52	22.95	34.12	3
DE-AT	Regression	—	0	0	**160.67**	**186.35**	2
	K-means: $k = 35$	35	30.20	31.07	177.03	215.71	3
	K-means: $k = 75$	75	17.41	17.92	183.17	210.04	2
	Fixed cell numbers: 9×9	69	36.82	36.92	182.63	210.67	3
	Fixed cell size: $1° \times 1°$	75	35.23	35.17	176.78	207.91	3

Table 3: Regression vs. classification approaches. *Median dist.* and *Mean dist.* are reported on the development set. *Eps.* refers to the number of fine-tuning epochs to reach minimum median distance.

diction as fine-tuning task. The corresponding results on the development sets are shown in Table 2. Our hypothesis was not verified, since for all three subtasks, the model with the large vocabulary yielded lower distances and converged faster than the model with the small vocabulary. We were not able to test other parameter settings due to time constraints.

4.2 Regression and classification approaches

The SMG task takes text as input and produces two outputs on a continuous scale, the predicted latitude and longitude. It is thus formulated most straightforwardly as a double regression task. However, three VarDial 2020 participants chose to convert the task into a classification task, grouping data points with similar coordinates into a discrete set of classes. We decided to replicate these three approaches on top of BERT and contrast them with our double regression approach.

K-means clustering Benites et al. (2020) apply k-means clustering to the VarDial training data, grouping instances with similar coordinates together. Each instance is then annotated with the centroid of its respective cluster. K-means clustering requires setting the parameter k, i.e. the number of clusters. Benites et al. (2020) choose a value of $k = 35$ for all subtasks on the basis of the CH development set. Besides $k = 35$, we also provide results for a second parameter choice, $k = 75$.

Grid with fixed number of cells Jauhiainen et al. (2020) lay a grid with 9 × 9 cells over the area of study. Each of the 100 corner coordinates of the 81 grid cells functions as an anchor point, and each instance is associated with its nearest anchor point. This approach yields thus a theoretical maximum of 100 labels, but not all of them are effectively assigned any instances.[2]

Grid with fixed cell size Hulden et al. (2015) also use a regular grid, but fix the size of its cells rather than the number of cells in the grid. Each instance is then associated with the center point of the grid cell it falls in. They experiment with roughly square grid cells of 0.5° × 0.5° to 10° × 10° on a different data set covering the USA. It is unknown which cell size was used in their VarDial 2020 submissions.[3] We decided to use a cell size of 0.2° × 0.2° for the CH subtask and of 1° × 1° for the BCMS and DE-AT subtasks.

[2] Note that Jauhiainen et al. (2020) introduce a second step in which they move the anchor points away from their initial (fixed) locations to the centroid of the data points assigned to them. With this second step, their approach can be conceived as a variant of k-means clustering with a specific initialization. For our experiments, we do not apply this second step.

[3] Note that Hulden et al. (2015) apply kernel density estimation after classification for smoothing. We did not implement this step to keep the approach comparable to the others.

The central part of Table 3 shows various statistics regarding the different approaches to space discretization. The number of effectively used classes is identical to the value of k for k-means clustering, but is typically a subset of the maximally available amount of grid cells, since the areas of study are not densely populated and do not fit into a square. The number of effectively used classes varied between 38 and 75.

We also computed the median reconstruction error, i.e., the median difference between the real coordinates and the coordinates of the assigned class labels. K-means turned out to provide lower errors than fixed-grid setups with comparable numbers of classes. This is again due to the uneven distribution of the SMG data points across space, most of which come from major cities and urban areas. For fixed-grid approaches, the reconstruction errors of the training and development set are expected to be similar, which is the case. For k-means, the development reconstruction error tends to be higher than the training one since the clusters were fitted to the training data alone. However, the k-means development reconstruction errors are still below the fixed-grid ones.

The right side of Table 3 contains the results of the experiments on the development data. In terms of distance, the (parameter-free) regression approach performs best on the CH and DE-AT subtasks, whereas the k-means approach with 75 clusters yields the best results for BCMS. The most likely reason for better performance of the classification approach on the BCMS data is high concentration of tweets in large cities, i.e., smaller dispersion of data than in the other two Jodel-based datasets. Classification approaches generally converged faster than regression. Among the different discretization approaches, no clear winner can be identified, despite the low reconstruction errors of k-means clustering.

4.3 Hyperparameter optimization

We performed hyperparameter optimization via the **wandb.ai** platform (Biewald, 2020), allowing for 30 iterations. We search for the optimal initial learning rate (we search between 9e-6 and 1e-4) and batch size (32, 64, 128, 256). Our experiments show that no significant improvements can be obtained by modifying the default hyperparameter values of the 5e-4 learning rate, setting with the batch size of 64 for all subtasks.

Pre-training	Median	Mean	Eps.
dbmdz + SwissCrawl	18.25	26.78	59
SwissCrawl (30k voc.)	22.10	31.16	10
SwissCrawl (3k voc.)	22.37	31.62	8

Table 4: Effect of different pre-training schemes for the CH unconstrained model, evaluated on the development set. *Eps.* refers to the number of fine-tuning epochs to reach minimum median distance.

4.4 Pre-trained models

Our unconstrained 2020 submissions showed that language-specific BERT models clearly outperformed multilingual BERT. For DE-AT, we continued using the DBMDZ model[4]. For the BCMS and CH subtasks, we carried out additional experiments.

BCMS While last year we achieved significant improvements by using the CroSloEngual BERT (Ulčar and Robnik-Šikonja, 2020)[5] in place of multilingual BERT, this year we had the opportunity to test a new model, BERTić (Ljubešić and Lauc, 2021).[6] Comparing the two models on the development data showed significant gains for the BERTić model, with a median distance of 30.11 in comparison to a median distance of 40.05 for the CroSloEngual BERT.

CH In the absence of a generally available pre-trained BERT model for Swiss German, our 2020 submission took the DBMDZ model as a starting point and continued pre-training for 10 epochs on the SwissCrawl corpus (Linder et al., 2020). We wanted to test whether pre-training (and defining the word segmentation model) on SwissCrawl alone would benefit classification performance. (SwissCrawl contains 562k sentences and 9.2M words, compared to 2350M words for DBMDZ.) Table 4 shows that this is not the case, neither with a 30k nor with a 3k vocabulary. Consequently, we stick to the 2020 setup for the final submission.

5 Final submissions

We present the final results of our constrained and unconstrained systems in Table 5. Unconstrained

[4] https://huggingface.co/dbmdz/bert-base-german-uncased

[5] https://huggingface.co/EMBEDDIA/crosloengual-bert

[6] https://huggingface.co/CLASSLA/bcms-bertic

Task	Setting	Development set		Test set	
		Median dist.	Mean dist.	Median dist.	Mean dist.
BCMS	Unconstrained	30.11	87.43	15.49	76.04
	Constrained	60.83	109.44	52.06	98.74
CH	Unconstrained	18.32	26.68	17.55	25.84
	Constrained	21.20	29.77	20.70	29.62
DE-AT	Unconstrained	147.88	172.04	149.33	172.52
	Constrained	160.03	184.53	161.13	184.97

Table 5: HeLju final submission scores.

systems, similar to last year, consistently outperform constrained systems. The results on the development and test set are consistent among all datasets, except for the BCMS task where our unconstrained system achieves a two times better score than on development data. The reason for this improvement is probably to be found in the way the 2021 data for the BCMS task was constructed. In this subtask, the test data consists of tweets that were published exclusively after March 2020, i.e., in times of COVID. It is quite likely that the limitation in overall mobility made the BCMS task significantly simpler.

6 Conclusion

In this paper, we presented our BERT-based approach to the problem of social media geolocation. This year we experimented primarily with tokenizer vocabulary size, space discretization and hyperparameter tuning. The hypothesis that a smaller tokenization vocabulary might improve generalizability and performance proved wrong. We showed that it actually performs worse and converges slower than if standard vocabulary size of 30k wordpieces is used. Converting the geolocation task to a classification task yielded worse results in all except the BCMS subtask. This is probably due to high concentration of BCMS Twitter data in the few large cities in the respective area. Hyperparameter tuning did not yield any consistent improvements and simply selecting the optimal epoch number on development data showed to be the best approach to this problem. Similarly to last year, unconstrained systems performed better than constrained systems. We share our code as well as the best-performing models via `https://github.com/clarinsi/geobert` and the HuggingFace models repository.

Acknowledgments

This work has been supported by the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 771113). This work has also been supported by the Slovenian Research Agency and the Flemish Research Foundation through the bilateral research project ARRS N6-0099 and FWO G070619N "The linguistic landscape of hate speech on social media", the Slovenian Research Agency research core funding No. P6-0411 "Language resources and technologies for Slovene language", and the European Union's Rights, Equality and Citizenship Programme (2014-2020) project IMSyPP (grant no. 875263).

References

Fernando Benites, Manuela Hürlimann, Pius von Däniken, and Mark Cieliebak. 2020. ZHAW-InIT - social media geolocation at VarDial 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 254–264, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Bharathi Raja Chakravarthi, Mihaela Găman, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial Evaluation Campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

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

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

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062, Mumbai, India. The COLING 2012 Organizing Committee.

Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. 2016. Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 213–217, Osaka, Japan. The COLING 2016 Organizing Committee.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

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

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2020. Experiments in language variety geolocation and dialect identification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 220–231, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Cristian Musat, and Andreas Fischer. 2020. Automatic creation of text corpora for low-resource languages from the internet: The case of Swiss German. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2706–2711, Marseille, France. European Language Resources Association.

Nikola Ljubešić and Davor Lauc. 2021. BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, Kiev, Ukraine. Association for Computational Linguistics.

Nikola Ljubešić, Tanja Samardžić, and Curdin Derungs. 2016. TweetGeo - a tool for collecting, processing and analysing geo-encoded linguistic data. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3412–3421, Osaka, Japan. The COLING 2016 Organizing Committee.

Yves Scherrer and Nikola Ljubešić. 2020. HeLju@VarDial 2020: Social media variety geolocation with BERT models. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 202–211, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Philippe Thomas and Leonhard Hennig. 2018. Twitter geolocation prediction using neural networks. In *Language Technologies for the Challenges of the Digital Age*, pages 248–255, Cham. Springer International Publishing.

Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: less is more in multilingual models.

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