Geolocation Prediction of Swiss Jodel Messages Using Large LMs

Karl Oskar Magnus Holm
Engineering Science and ICT - Department of Geomatics / 2023
koholm@stud.ntnu.no

Abstract

This paper takes on the shared task on Social Media Variety Geolocation (SMG) from the VarDial workshops in 2020 and 2021, focusing on the subtask of predicting geolocations of Swiss Jodel messages. The winner of both year's competitions used BERT Transformer models, and this project builds upon their work, investigating if newer language-specific models, other map projections, or different hyperparameters can improve accuracy. While models trained in this project were unable to surpass the best results from 2020 and 2021, one can deduce from the results that language-specific models perform best, and that metric map projections are the preferred way of representing coordinates for the task at hand. Language-specific variants of Google's BERT and Meta's X-Mod were tested, with the former achieving by far the best results.

1 Introduction

This project is based on the shared task on Social Media Variety Geolocation (SMG) from VarDial 2020 and 2021 (seventh and eighth editions, respectively), the Workshop on Natural Language Processing (NLP) for Similar Languages, Varieties, and Dialects (Chakravarthi et al., 2021; Gaman et al., 2020). The aim of the task differs somewhat from the most common types of NLP VarDial tasks, where the goal typically is to choose from a finite set of variety labels (Scherrer and Ljubešić, 2021, p. 1). Here, the goal is to predict a set of scalars, namely the latitude and longitude from which a social media post was posted. This VarDial task stayed the same from 2020 to 2021, including three language areas: the Bosnian-Croatian-Montenegrin-Serbian language area, the German language area (Germany and Austria), and the German-speaking Switzerland.

This project is limited to the latter of these language areas, namely the German-speaking Switzerland. Reasons for this include the limited time scope of the task and having to share the necessary computing resources with fellow students at the department. The goal is to try and recreate the results of Scherrer and Ljubešić (2020), who used a BERT-based classifier, viewing the problem as a double regression task. I focus on the 2020 dataset because there were a lot more submissions this year compared to 2021, where there was little time between the announcement of the shared task and the submission deadline (Chakravarthi et al., 2021, p. 6).

2 Background

This section will elaborate on technologies central to this project. It is assumed that the reader has a basic understanding of what Language Models (LMs) are, and also that they are somewhat versed in the world of Natural Language Processing (NLP).

2.1 The Transformer Architecture

Vaswani et al. (2017) managed to achieve new state-of-the-art results for machine translation tasks with their introduction of the Transformer architecture. The Transformer has later been proved effective for numerous downstream tasks, and for a variety of modalities. Titling their paper 'Attention Is All You

Need', Vaswani et al. suggest that their attention-based architecture renders Recurrent Neural Networks (RNNs) redundant, due to its superior parallelization abilities and the shorter path between combinations of position input and output sequences, making it easier to learn long-range dependencies (Vaswani et al., 2017, p. 6).

The Transformer employs self-attention, which enables the model to draw connections between arbitrary parts of a given sequence, bypassing the long-range dependency issue commonly found with RNNs. An attention function maps a query and a set of key-value pairs to an output, calculating the compatibility between a query and a corresponding key (Vaswani et al., 2017, p. 3). Looking at Vaswani et al.'s proposed attention function (1), we observe that we take the dot product between the query Q and the keys K, where Q is the token that we want to compare all the keys to. Keys similar to Q will get a higher score, e.g., be *more attended to*. These differences in attention are further emphasized by applying the softmax function. The final matrix multiplication with the values V, being the initial embeddings of all input tokens, will give us a new embedding in which all tokens have some context from all other words. We improve the attention mechanism by multiplying queries, keys, and values with weight matrices learned through backpropagation. Self-attention is a special kind of attention in which queries, keys, and values are all the sequence.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{1}$$

Attention blocks can be found in three places in the Transformer architecture (Vaswani et al., 2017, p. 5) (I will use machine translation from Norwegian to German as an example):

1. In the encoder block to perform self-attention on the input sequence (which is in Norwegian)
2. In the decoder block to perform self-attention on the output sequence (which is in German)
3. In the decoder block to perform cross-attention (or encoder-decoder attention) where each position in the decoder attends to all positions in the encoder

The Transformer represented a breakthrough in the field of NLP, and is the fundamental building block of LMs like BERT.

2.2 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a family of language models which was first introduced in 2018 and is designed to facilitate a wide range of downstream tasks (Devlin et al., 2019, May, p. 5). The BERT architecture consists of stacked bidirectional Transformer encoders. The self-attention mechanism allows for training of deep bidirectional representations. The input sequence is transformed into embeddings (vector representations). These per-token embeddings include information about the meaning of the word itself, the meaning of the sentence/segment it belongs to, and the token's position in the full input. These embeddings then pass through a stack of Transformer encoders (12 and 24 for $\textbf{BERT}_{\textbf{BASE}}$ and $\textbf{BERT}_{\textbf{LARGE}}$, respectively), allowing the model to learn more complex patterns and of different granularities (token, sentence, document) (Devlin et al., 2019, May, p. 5).

In the BERT framework, there are two training steps, namely the pre-training and fine-tuning procedures. BERT is pre-trained on two NLP tasks. One is Masked Language Modelling (MLM), in which 15% of words are masked with the special `[MASK]` token and are left for the model to predict (Devlin et al., 2019, May, p. 4). The MLM task helps the model learn bidirectional representations. The second of the two unsupervised tasks used during pre-training is Next Sentence Prediction (NSP), where the special `[CLS]` token (found at the start of each tokenized sequence) is used to predict if a sentence `B` follows `A`. During this pre-training step, the input sequence looks like this:

`[CLS]` this is sentence A `[SEP]` and this is sentence B `[SEP]`

The `[CLS]` token is used to label sentence B as either `IsNext` or `NotNext`.

BERT is normally fine-tuned to specific downstream tasks by using the `[CLS]` token, which captures an aggregated representation of the input sequence. This vector representation can then be used as input to a classification layer for tasks like multi-label classification and regression.

2.3 X-Mod

Cross-lingual Modular (X-Mod) models (Pfeiffer et al., 2022, July) attempt to tackle the common problem of multilinguality in language models. Typically, when one attempts to train a language model to be multilingual by training on numerous languages, the performance tends to drop after reaching a certain level - *the curse of multilinguality*. The model creators at Meta AI claim that X-Mod mitigates the negative interference between languages and unlocks improved monolingual and cross-lingual performance (Pfeiffer et al., 2022, July, p. 1).

2.4 Geodesic Terminology and Metrics

The evaluation is based upon the Haversine formula (2), with the Earth's radius assumed to be 6371 km. The evaluation metric is the median Haversine distance between the predicted coordinates and the ground truth (Scherrer and Ljubešić, 2020, p. 4). A formulation of the Haversine distance can be found on its Wikipedia page where it is described as "the great-circle distance between two points on a sphere given their longitudes and latitudes". The distance d can be expressed as

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{2}$$

where ϕ and λ are latitude and longitude values.

Different map projections were used in the project. The Universal Transverse Mercator (UTM) map projection splits the Earth's surface into 60 zones in the latitudinal direction and 19 zones in the longitudinal direction, forming a grid. This allows for expressing coordinates in meters within a grid zone while still obtaining high accuracy measurements. Coordinate values in UTM are typically in the six-figure range, with the easting of the central meridian[1] defined as 500,000 meters to avoid negative easting values within the zone.

The Swiss coordinate system, LV95 (Federal Office of Topography swisstopo, n.d.), was also explored. Its fundamental reference point is set to the Swiss capital of Bern, and the values typically lie in the seven-figure range.

3 Related Work

This paper builds upon the work of Scherrer and Ljubešić (2020). They were the only participants in VarDial who used large LMs like BERT in the shared task on Social Media Variety Geolocation (SMG), and they did so with great success, winning the shared task in both 2020 and 2021. Scherrer and Ljubešić converted the task into a double regression problem, where they predicted latitude and longitude from the output `[CLS]` representation of BERT models. They experimented with different pre-trained models, coordinate encodings, and hyperparameters. Their main finding was that single-language models outperform multilingual models, which perform worse due to capacity dilution and tokenizers yielding suboptimal text splitting (Scherrer and Ljubešić, 2020, p. 3). As they were unable to find a pre-trained model for Swiss German, they instead trained `bert-base-german-uncased`[2] (German BERT) on the SwissCrawl corpus (Linder et al., 2020, June). Training a total of 48 models, Scherrer and Ljubešić achieved a median distance of 15.72 km in this unconstrained setting using the default data split. They reduced the median distance to 15.45 km by using a substantial portion of the development set for training (Scherrer and Ljubešić, 2020, p. 6).

Gaman et al. (2020) and Chakravarthi et al. (2021) summarize the findings in the 2020 and 2021 editions of VarDial, including attempts made on the SMG task. While Scherrer and Ljubešić (2020)

[1] https://gisgeography.com/central-meridian/

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

	lat	lon	text
0	47.22	7.43	Dr Chester Bennington isch tot (pensive face)(pensive face)(pensive face) #rip #linkinpark Dr Manager heds bestätiged (expressionless)...
1	46.86	8.21	Mini Fründin hed Lust uf Doktorspieli gha... ... sie hocked jetzt sit 2 Stund...
2	47.39	8.18	Slayer isch besser. Det han ich gescht mini Drohne stiege lah (smiling face with smiling eyes) Cool was hesch f...

Table 1: The first three rows of the training dataset

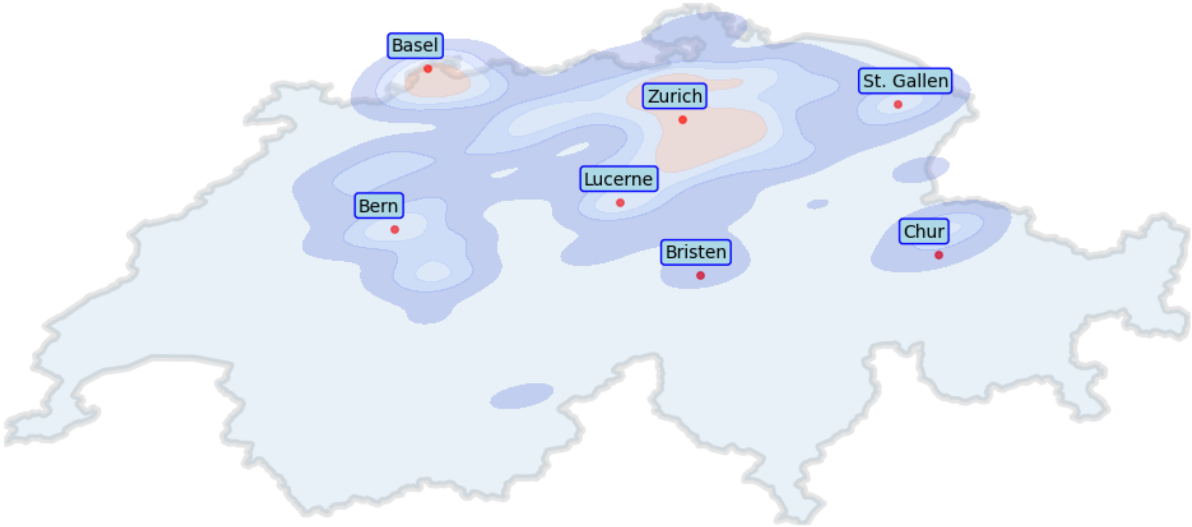

Figure 1: Heatmap of the training data

generally dominated the leaderboards, Benites de Azevedo e Souza et al. (2020) proposed a method that performed best among constrained submissions on the Swiss task (Gaman et al., 2020, pp. 8–9), and was only marginally worse than Scherrer and Ljubešić's unconstrained submissions. Benites de Azevedo e Souza et al. (2020) used K-Means clustering (Lloyd, 1982) of locations and predicted cluster identities, framing the problem as a classification task rather than a regression task. Their best submission extracted features from different levels of token granularity, training a separate Support Vector Machine (SVM) for each feature set, before feeding the distances to the decision boundaries for each feature classifier as input to a SVM meta-classifier.

4 Datasets

Data from VarDial 2020 and 2021 was acquired from a GitHub repository created by Yves Scherrer, co-author of the winning solution for the SMG task in both 2020 and 2021. All but one dataset have a ground truth associated with them, and I assume this unlabelled dataset was used for a private leaderboard. Three labelled datasets with 22,600, 3,097, and 3,068 samples were provided. They serve for training, development/validation, and testing, respectively. Table 1 shows the first three rows of the training dataset. It was collected by Hovy and Purschke (2018, October, pp. 2–3) using the (at the time) publicly available Jodel API.

While Switzerland is a country with four official languages (Swiss-German, French, Italian, and Romansh Grishun), the dataset contains only Swiss-German Jodel messages, presumably because the task focuses on dialectal differences. This results in the heatmap in Figure 1.

Attempts were made to acquire a Norwegian dataset based on Norwegian Twitter/X messages using

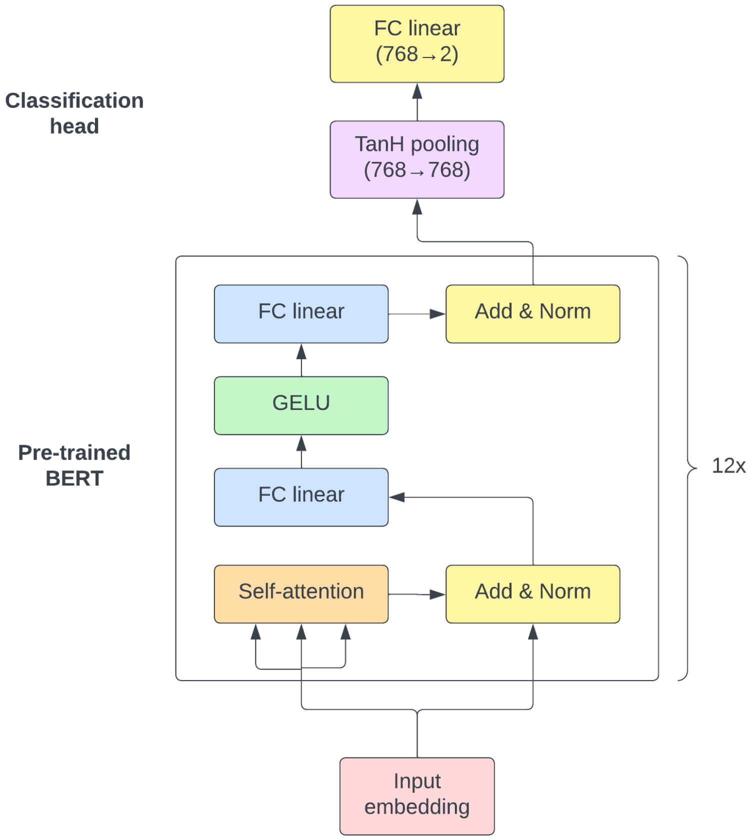

Figure 2: Model architecture

the method described in Ljubešić et al. (2016), but due to recent changes in the Twitter/X API[3], these attepmts failed. The possibility of using Norwegian Jodel messages was also explored, but to the author's knowledge, their API is no longer available to the public.

5 Model

Figure 2 shows a rough model architecture. It consists of a pre-trained BERT model with a classification head on top. This architecture is not representative for the X-Mod-based model. The classification head has two outputs, namely latitude and longitude coordinates. It takes as input the output corresponding to the `[CLS]` token, which captures the aggregated sequence representation. The hyperbolic tangent activation function adds some non-linearity to the output before it is fed into a fully connected linear layer, where latitude and longitude values are predicted.

The best results of Scherrer and Ljubešić (2020) came from using a language-specific BERT-based model. As no pre-trained model was found in 2020, they pre-trained the `bert-base-german-uncased` model on the SwissCrawl corpus (Scherrer and Ljubešić, 2020, pp. 3–4). Since then, models pre-trained on Swiss corpora have emerged. Three different pre-trained models were tested in this project. These are `bert-base-german-uncased`[4], `bert-base-german-cased-finetuned-swiss`[5], and `ZurichNLP/swissbert`[6]. `bert-base-german-uncased` is a German-language model pre-trained on a Wikipedia dump, the EU Bookshop corpus, and more. `bert-base-german-cased-finetuned-swiss` is based on `bert-base-german-cased` and is pre-trained on the Leipzig Corpora Collection (Goldhahn et al., n.d.) and SwissCrawl (Linder et al., 2020, June).

[3] https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research
[4] https://huggingface.co/dbmdz/bert-base-german-uncased
[5] https://huggingface.co/statworx/bert-base-german-cased-finetuned-swiss
[6] https://huggingface.co/ZurichNLP/swissbert

ZurichNLP/swissbert is the only non-BERT model used. It is based on X-Mod, with adapters trained for German, French, Italian, and Romansh Grishun.

Pre-trained models where accessed through Huggingface's transformers library.

6 Experiments and Results

In this section, I will elaborate on my approach to training models, before presenting the most important findings from my experiments.

6.1 Experimental Setup

All experiments were performed on an NVIDIA GeForce RTX 4090 with 24 GB G6X memory[7]. Computing resources belong to the Department of Geomatics at NTNU and are shared with fellow 5th year geomatics students. PyTorch[8] was used to create a training loop, and Huggingface's transformers library was used to fetch pre-trained models from their hub.

Because of the limited timespan and computing resources of this project, I opted to freeze some of Scherrer and Ljubešić's hyperparameters. This includes the maximum sequence length and the batch size (128 and 32, respectively), the latter of which was also limited by the GPU memory. The loss function (MAE/L1) and the joint scaler (Scherrer and Ljubešić, 2020, p. 5) also remained largely unchanged. The focus of my experiments was to compare the performance of different pre-trained models and to see what effect different coordinate projections and learning rates have.

6.2 Experimental Results

A total of 15 models were trained. Table 2 shows a selection of the most interesting results along with the map projection and learning rate/scheduler used. A full overview of configurations and their corresponding results can be found on the project's GitHub page. Fine-tuned models and training logs can be found in this Google Drive folder.

The best results were achieved when using the statworx/bert-base-german-cased-finetuned-swiss pre-trained model with a reduced development/validation dataset, using 1000 samples for validation and the rest for training. Learning rate schedulers did not prove very efficient for this task, and a low learning rate of 2e-5 yielded the best results instead. No real difference in performance was observed when using the UTM projection over raw latitude/longitude values, but the Swiss-native LV95 projection gave the model a performance boost into the 15-kilometer range. The X-Mod based ZurichNLP/swissbert model did not prove efficient for the task at hand.

Figure 3 shows the pointwise distance from the ground truth when using the best-performing model to make predictions on the test gold dataset.

7 Discussion

Results show that the Swiss LV95 map projection proved most efficient for predicting geolocation from Swiss Jodel messages. This may indicate that using a metric Coordinate Reference System (CRS) like LV95 over a spherical representation like latitude and longitude values can be beneficial in a double regression task for predicting geographical coordinates. It could seem that the model finds it harder to learn spherical representations. These findings are counter to those of Scherrer and Ljubešić (2020, p. 5), who found that raw latitude and longitude values do not perform worse than metric projections. They only did tests on the UTM projection, however, and did not use the LV95 projection.

Not surprisingly, the language-specific model (bert-base-german-cased-finetuned-swiss) proved most suitable for this task. Being pre-trained on large Swiss corpora, its creators were able to show a 5 percent improvement over its German parent model. It seems this pre-training enhances the model's ability to pick up on dialectal details in the data. The X-Mod-based swissbert model, which is based on a model designed to be multilingual (Pfeiffer et al., 2022, July), did not seem to

[7] https://www.nvidia.com/nb-no/geforce/graphics-cards/40-series/rtx-4090/

[8] https://pytorch.org/

Pre-trained model	Coordinate Projection	LR/Scheduler	Median Distance [km]
dbmdz/**bert-base-german-uncased**	lat/lon	4e-5	17.81
statworx/**bert-base-german-cased-finetuned-swiss**	lat/lon	4e-5	17.08
statworx/**bert-base-german-cased-finetuned-swiss**	UTM	ReduceLROnPlateau	16.52*
statworx/**bert-base-german-cased-finetuned-swiss**	UTM	2e-5	16.05*
statworx/**bert-base-german-cased-finetuned-swiss**	lat/lon	2e-5	16.19*
statworx/**bert-base-german-cased-finetuned-swiss**	LV95	2e-5	**15.76***
ZurichNLP/**swissbert**	UTM	2e-5	17.59*

Table 2: Highlighted results

* Proportion of developmentset used as additional samples for training

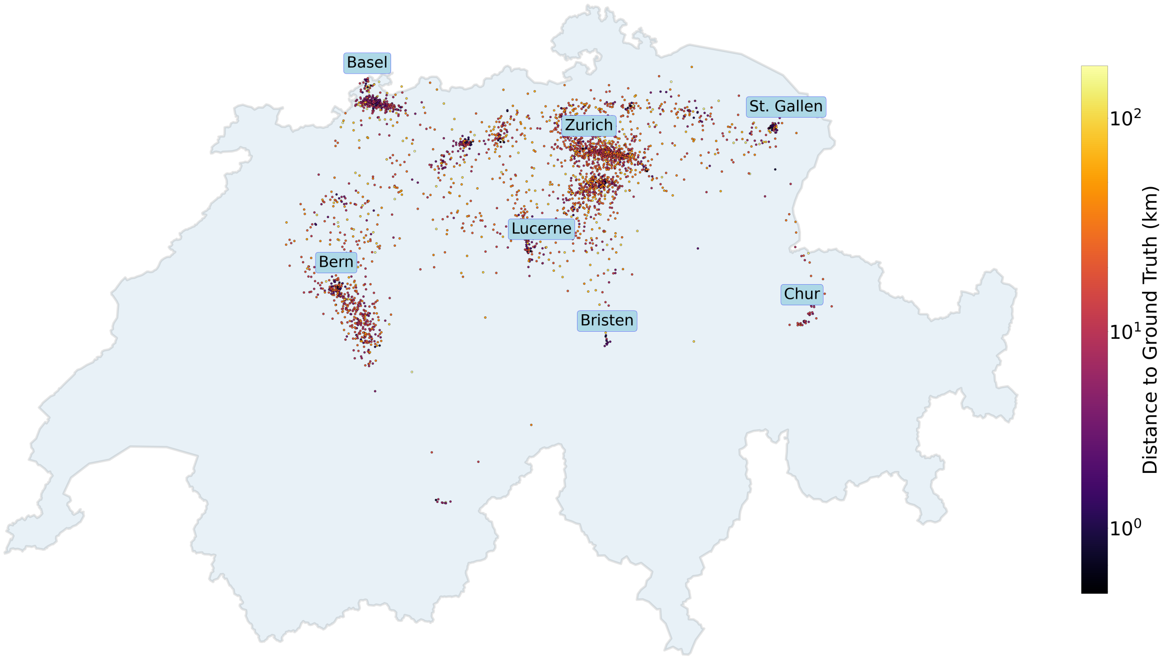

Figure 3: Pointwise distance from the ground truth

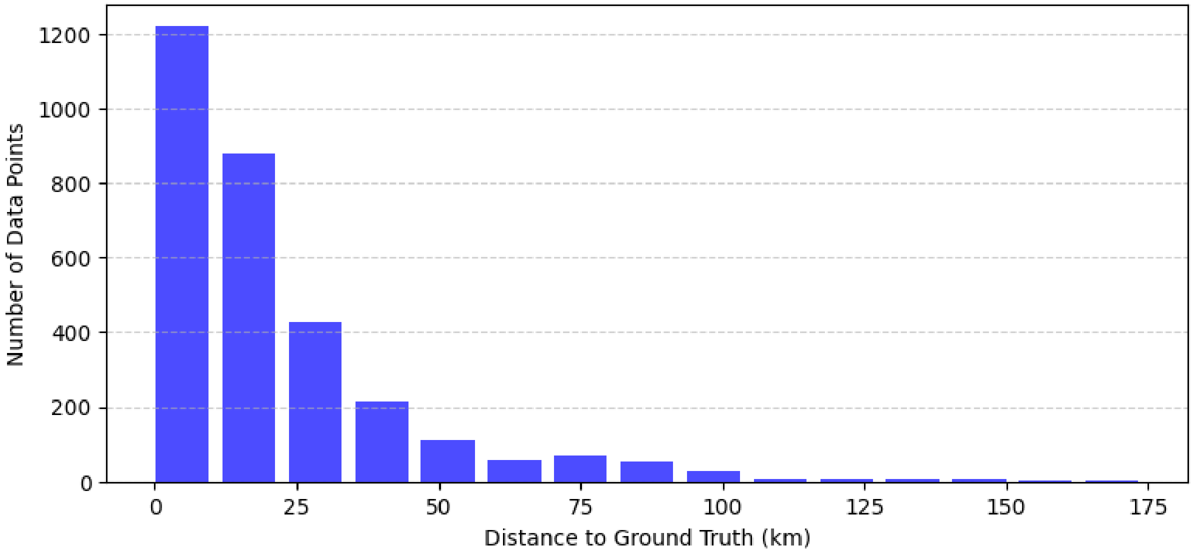

Figure 4: Error distribution: distances from predicted locations to ground truth

possess the same dialectal knowledge and performed only marginally better than the German `bert-base-german-uncased` model.

Furthermore, it is clear from the results that the learning rate schedulers used did not improve the test score. The `ReduceLROnPlateau` and `OneCycleLR` schedulers were tested, and while they greatly reduced the convergence time, they were unable to achieve satisfactory median distances. Since the schedulers showed such little promise for this task, they were not investigated much further. I do think, however, that they could prove efficient if one can find a suitable set of initialization parameters.

Overall, the results are quite good and would have sufficed for a second place in the VarDial 2020 competition. One would expect, however, that with newer models like `bert-base-german-cased-finetuned-swiss`, one should be able to achieve better results while using methods similar to those used in 2020. Pinpointing an exact reason why this did not happen is difficult, but Scherrer and Ljubešić (2020) having trained a total of 48 models as opposed to the 15 of this study could be one of them. There may also be some default hyperparameters in the `simpletransformers` library that Scherrer and Ljubešić (2020) did not discuss in their paper which made it difficult to recreate their results, or there might be other totally different issues with the implementation in this project.

8 Conclusion and Future Work

This paper addressed the shared task on Social Media Variety Geolocation from VarDial 2020 and 2021, and shows that newer language-specific Language Models are very capable of predicting latitude/longitude pairs from text samples, obtaining a median error of 15.76 kilometers on the Swiss task. It also suggests that using metric Coordinate Reference Systems could be better for what becomes a double regression problem. The best model achieved results which would have sufficed for a second place in the 2020 competition.

Further research should look into using a classifier based upon the $\textbf{BERT}_{\textbf{LARGE}}$ base model. As of 26th November 2023, there is no such model pre-trained on Swiss corpora known to the author of this paper. Having twice as many encoder layers as the $\textbf{BERT}_{\textbf{BASE}}$ model, it should be able to pick up on finer dialectal details, if pre-trained.

References

Benites de Azevedo e Souza, F., Hürlimann, M., von Däniken, P., & Cieliebak, M. (2020). ZHAW-InIT : Social media geolocation at VarDial 2020. *Workshop on NLP for Similar Languages, Varieties and Dialects, Barcelona (Spain), Online, 13 December 2020*, 254–264
Accepted: 2021-02-04T13:13:06Z.

Chakravarthi, B. R., Mihaela, G., Ionescu, R. T., Jauhiainen, H., Jauhiainen, T., Lindén, K., Ljubešić, N., Partanen, N., Priyadharshini, R., Purschke, C., Rajagopal, E., Scherrer, Y., & Zampieri, M. (2021). Findings of the VarDial Evaluation Campaign 2021. *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, 1–11.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Federal Office of Topography swisstopo. (n.d.). The Swiss coordinates system.

Gaman, M., Hovy, D., Ionescu, R. T., Jauhiainen, H., Jauhiainen, T., Lindén, K., Ljubešić, N., Partanen, N., Purschke, C., Scherrer, Y., & Zampieri, M. (2020). A Report on the VarDial Evaluation Campaign 2020. *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 1–14.

Goldhahn, D., Eckart, T., & Quasthoff, U. (n.d.). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages.

Hovy, D., & Purschke, C. (2018, October). Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting. In E. Riloff, D. Chiang, J. Hockenmaier & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4383–4394). Association for Computational Linguistics.

Linder, L., Jungo, M., Hennebert, J., Musat, C., & Fischer, A. (2020, June). Automatic Creation of Text Corpora for Low-Resource Languages from the Internet: The Case of Swiss German.

Ljubešić, N., Samardžić, T., & Derungs, C. (2016). TweetGeo - A Tool for Collecting, Processing and Analysing Geo-encoded Linguistic Data. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3412–3421.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137.

Pfeiffer, J., Goyal, N., Lin, X., Li, X., Cross, J., Riedel, S., & Artetxe, M. (2022, July). Lifting the Curse of Multilinguality by Pre-training Modular Transformers. In M. Carpuat, M.-C. de Marneffe & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3479–3495). Association for Computational Linguistics.

Scherrer, Y., & Ljubešić, N. (2020). HeLju@VarDial 2020: Social Media Variety Geolocation with BERT Models. *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 202–211.

Scherrer, Y., & Ljubešić, N. (2021). Social Media Variety Geolocation with geoBERT. *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, 135–140.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*.