Geolocation Prediction from Jodel and Twitter Messages

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

Abstract

1 Introduction

This project is based on the shared task on Social Media Variety Geolocation (SMG) from VarDial 2020 and 2021 (seventh and eighth edition, 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 in this case), and the German-speaking Switzerland.

This project is limited to the latter of these language areas, that is, 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, making it a double regression task. I focus on the 2020 dataset because there were a lot more submissions this year as opposed to in 2021, due to the short time between the announcement of the shared task and the submission deadline (Chakravarthi et al., 2021, p. 6).

The reason for picking the task on the German-speaking Switzerland is its similarities to the dialectal landscape of Norway. Røyneland (2009, p. 14) writes that "dialects in Norway have had and still have a much stronger position than dialects ... in most of Europe, with the exception of the German-speaking part of Switzerland.". I therefore find it reasonable to assume that a method which works well on the Swiss dataset could also perform well on a Norwegian dataset. Unfortunately, I was unable to find a suitable Norwegian dataset, and creating one proved too difficult and time-consuming to be worthwhile, considering the time horizon of the project.

2 Background

This section will elaborate on technologies central to this project. It is assumed that the reader has basic understanding of what Language Models (LMs) are, and also that they are somewhat wandered 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. Titleing 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 parallellization 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, by-passing 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 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 is further emphasized be applying the softmax function. The final matrix multiplication with the values V, being the intitial 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 mutiplying each query, key, and value with weigth matrices learned through backpropagation. Self-attention is a special kind of attention in which queries, keys, and values are all the same input 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 (Vaswani et al., 2017, p. 5) in the Transformer architecture (I will use machine translation as example, say, from Norwegian to German):

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 established a new state of the art in machine translation Vaswani et al. (2017), 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 (Devlin et al., 2019, May). BERT is designed to facilitate a wide range of downstream tasks. The self-attention mechanism allows for efficient bidirectional cross attention between two sentences in one step, a process which previously was done by encoding text pairs before applying bidirection cross attention (Devlin et al., 2019, May, p. 5). See Figure 1.

2.3

2.4 Geodesic Terminology and Metrics

Haversine formula, etc.

2.5 (Double) Regression

BERT can be used for regression tasks.

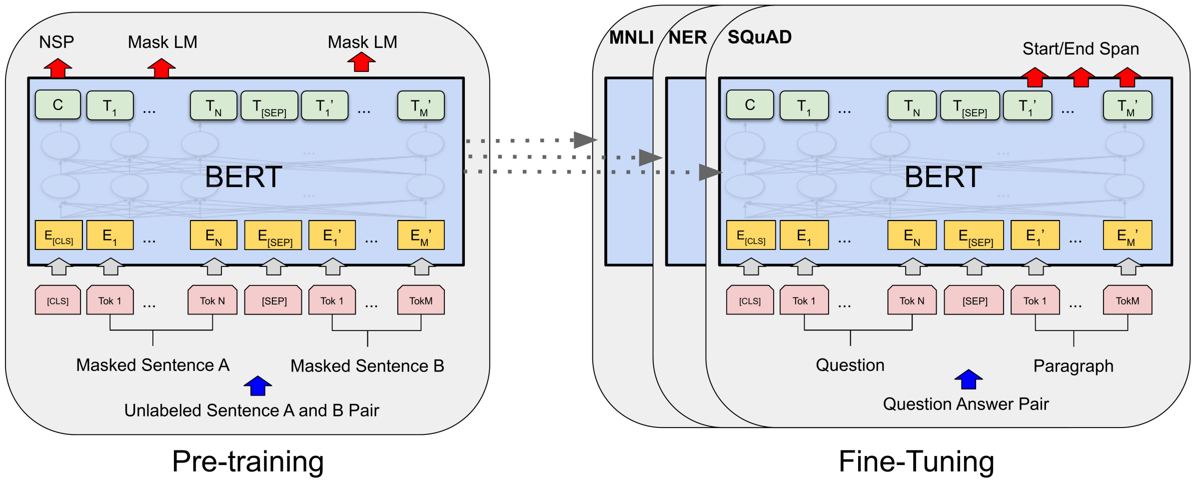

Figure 1: High-level overview of the pre-training and fine-tuning procedures for BERT (Devlin et al., 2019, May, p. 3).

3 Related Work

4 Datasets

5 Model

6 Experiments and Results

6.1 Experimental Setup

All experiments were performed on an NVIDIA GeForce RTX 4090, having 24 GB G6X memory.[1] Computing resources belong to the Department of Geomatics at NTNU and are shared with fellow 5th year geomatics students. I opted for PyTorch[2] in my experiments unlike Scherrer and Ljubešić (2020), who used the high-level SimpleTransformers[3] library.

6.1.1 German-speaking Switzerland

Data from VarDial 2020 and 2021 was acquired from a GitHub repository[4] created by Yves Scherrer's, co-author of the winning solution for the SMG task, both years.

The best results of Scherrer and Ljubešić (2020) came from using a language-specific BERT. As no pre-trained model was found in 2020, they fine-tuned the pre-trained `bert-base-german-uncased`[5] model on the SwissCrawl corpus (Scherrer and Ljubešić, 2020, pp. 3–4). Since then, a pre-trained Swiss BERT model has been released[6], which I used in my experiments.

6.1.2 Norway

6.2 Experimental Results

7 Evaluation and Discussion

8 Conclusion and Future Work

The code and data used during model development is available at `https://github.com/oskarhlm/TDT13`.

[1] https://www.nvidia.com/nb-no/geforce/graphics-cards/40-series/rtx-4090/
[2] https://pytorch.org/
[3] https://simpletransformers.ai/
[4] https://github.com/yvesscherrer/vardial-shared-tasks
[5] https://huggingface.co/bert-base-german-uncased
[6] https://huggingface.co/statworx/bert-base-german-cased-finetuned-swiss

References

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.

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.

Røyneland, U. (2009). Dialects in Norway: Catching up with the rest of Europe? *2009*(196-197), 7–30.

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*.