Mixture of Experts for Translating Different Aromanian Dialects

1st Semester of 2023-2024

Gheorghe Nistor
gheorghe.nistor@s.unibuc.ro

Dan-Andrei Ungureanu
dan-andrei.ungureanu@s.unibuc.ro

Abstract

In this project we aim to use the Mixture of Experts machine-learning technique in constructing an AI Model capable of translating dialects of the Aromanian language to Romanian. We have 3 separate dialects, and we have trained a separate model for each. We use a gating function to identify the dialect of a given text, in order to select the suitable model ('expert') to be used in processing the translation.

1 Introduction

For millennia, language translation has been a key component of inter-human and inter-civilization interactions. On the one hand, it is important for many day-to-day situations, such as understanding other people in a conversation, translating a document, reading instructions and directions, if, for example, you are travelling in a foreign country. It is also a vital tool in the study of history, where, very often, our knowledge is limited by the capacity to read and understand old written sources, the language in which they are written being long forgotten.

Many such texts are lost to history due to the inability to reconstruct their corresponding languages. Two principal causes of this are the lack of active speakers, and the scarcity of texts to serve as a sufficient base for usual language reconstruction methods. While these two problems constitute a solid hindrance to any attempt to decipher an old language, it does not and should not stop us from trying to preserve the heritage of those languages that we still know enough about, by recording and studying them.

For example, the decline of the Irish language has seen a lot of media attention in the 2010s, mainly due to Ireland's complicated history and many foreign attempts to suppress Irish cultural identity throughout the centuries. As of 2022, roughly 2 million people still spoke the language, but only 200.000 reported to speak it 'very well'. The numbers do not seem to indicate an endangered language, considering some have less than 100 active speakers, but it is still a rather low percentage for a total of 70 million people who claim Irish descent.

Another example, which we have tried to tackle in this project, is the Aromanian language and its dialects. Aromanian is a romance language spoken mostly in the Balkans, thought to have developed from proto-Romanian, in a split that happened around the 10th Century AD. It has several dialects, each spoken in little 'enclaves' of territory, surrounded and influenced by larger language groups (Greek, Slavic). It is generally associated with the 'Vlachs' people, speakers of romance languages living South of the Danube. Aromanian has several dialects, split into two categories by phonetic comparison: North-Western (North Macedonia, Albania) and South-Eastern (Greece). Around 200.000 people are reported to speak Aromanian, and the language only has official status in Northern Macedonia (Albania, however, does recognise the people as an official minority).

With the rich cultural environment in the aforementioned regions, it is indisputable that the preservation of this language and its dialects is a major addition to our world heritage, and a significant step in the understanding of local history.

To this end, we worked on several models that can translate from these dialects to Romanian, the most closely-related 'parent' language of Aromanian. However, this still leaves the problem of correctly identifying the dialect we wish to translate from, which could prove significantly difficult for the average user. Hence, we introduce the Mixture of Experts machine-learning technique, creating a larger model that makes use of the individual ones we created before. Conceptually, the Mixture of Experts will first attempt to recognise the problem at hand, and then delegate the task to the most suit-

able model to handle it and produce an output (it is possible for the task to be assigned to multiple potentially suitable models, but in this project we only use one). We will first attempt to identify the dialect in which the to-be-translated text is written, and then assign it to the afferent pre-trained model to translate it to the desired language (in this case, Romanian).

We split the assignment into the following tasks, by member:

- Gheorghe Nistor - model training, gating function
- Dan Ungureanu - documentation, text normalization, building dataframe

We worked together on writing most of the code as the tasks were interconnected and it would have proven much more difficult to work individually and then attempt to merge what we built.

2 Approach

The first step was designing our Mixture of Experts, which was pretty straightforward. We would have a separate model for each dialect, capable of translating it into Romanian, and a gating function which would identify a given text's language and select the appropriate model to give us the translation.

The following step , beginning the implementation, was building a dataframe from the training data sets, enabling us to train the gating function into correctly recognising the language of the input, and the translation models themselves. In this step we also normalized the data - the dialects contained several non-standard characters, corresponding to some orthographic and phonetic particularities, which we had to convert into UTF-8 characters. For convenience, we also converted all the texts to lower-case.

In the next step we handled translation. For this part, we used a scaled-down version of Google's T5 transformer (t5-small) specialized on natural language processing. We refined it for handling our particular tasks (translating from a dialect to Romanian).

We then used a logistic regression model to enable us to determine a given text's language, to then appropriately select the corresponding translation model, effectively implementing the Mixture of Experts technique.

Overall, we have 3 separate scripts that implement our model. The *aromanian_text_cleaning.ipynb* script is responsible for preprocessing the data. It builds the dataframe, replaces non-standard characters and transforms all texts to lowercase. We use the same procedure later on in the translation script for processing the input text (the text to be translated).

The *aromanian_translation_model_training.ipynb* script covers the training of the translation models, by fine-tuning the T5 model.

The *aromanian_translation_execution.ipynb* script leverages the models generated in the previous step for executing the translation tasks. It includes a logistic regression model for language detection to ensure the correct translation model is applied based on the input text's language.

For further references see the Github repo

3 Limitations

The main limitation of our model is the ability to distinguish between languages when the choice for the most suitable translation model needs to be made. It is most apparent in the case of dialects, as they are quite similar and this produces confusion.

Accuracy: 0.8681925808997633

	precision	recall	f1-score	support
en	1.00	0.97	0.99	418
es	0.96	0.98	0.97	436
fr	0.99	0.97	0.98	375
rup	0.64	0.73	0.68	421
rup_cun	0.94	0.93	0.94	431
rup_std	0.73	0.64	0.68	453
accuracy			0.87	2534
macro avg	0.87	0.87	0.87	2534
weighted avg	0.87	0.87	0.87	2534

As can be seen in the above report, the model is having a difficult time distinguishing between the rup and rup-std dialects.

In terms of scalability, the Mixture of Experts approach is, perhaps, the most suitable for the task at hand, as it can be easily enhanced with new translation options.

4 Conclusions and Future Work

In conclusion, we have successfully achieved our objective of constructing a MoE to handle the translation of Aromanian dialects to Romanian. In terms of improvement, we could use a more heavy-weight model/tokenizer for the translation step, which would improve the quality of the translations at the cost of more computing time, but for simplicity we stuck to t5-small.

It was definitely an excellent learning opportunity as Mixture of Experts is a highly versatile and

configurable technique, and we might end up using it in future Machine-Learning projects.

References

Mixture of Experts

- https://huggingface.co/blog/moe
- https://machinelearningmastery.com/mixture-of-experts/
- Jordan, M. I., Jacobs, R. A. (1994). Hierarchical Mixtures of Experts and the EM Algorithm. Neural Computation, 6(2), 181–214. doi:10.1162/neco.1994.6.2.181
- https://blog.research.google/2022/11/mixture-of-experts-with-expert-choice.html?m=1

T5 model - https://huggingface.co/docs/transformers/en/model_doc/t5

Census 2022: Number of Irish speakers increases but only 10% can speak it very well - Irish Language Census

Aromanian language - Wikipedia Aromanian Language