One-shot Prompt for Language Variety Identification

Nat Gillin
nat.gillin@gmail.com

Abstract

We present a one-shot prompting approach to multi-class classification for similar language identification with an off-the-shelf pre-trained large language model that is not particularly trained or tuned for the language identification task. Without post-training or fine-tuning the model, we simply include one example per class when prompting the model and surprisingly the model is able to generate the language and locale labels accordingly.

1 Introduction

Recent works validated the idea of using language models generation performs well in classification task (Li et al., 2018a; Thant and Nwet, 2020; Hadar and Shmueli, 2021) and generation models can also perform competitively as zero-shot text-classifiers (Yin et al., 2019; Meng et al., 2022; Sun et al., 2023; Wang et al., 2023). Particular to language identification, Gillin (2022) have trained an encoder-decoder model for the French Cross-Domain Dialect Identification (FDI) dataset (Găman et al., 2023) for the VarDial 2022 shared task (Aepli et al., 2022a).[1]

Previously one might find it appealing to train or fine-tune a model to achieve state-of-the-art NLP models for specific tasks, but recent advancement in large language models and prompt-based solutions have made us think,

> *What if we just prompt a popular LLM and make it work like a classifier without tuning it?*

To test the idea of just prompting a pre-trained model for language identification, we evaluated the approach on the *English*, *French*, *Portuguese* and *Spanish* subset of the DSL Multi-label classification of similar languages (DSL-ML) shared task at VarDial 2024 (Chifu et al., 2024).[2] A few of example inputs and outputs of the DSL-ML test data are as follows:

> **[IN]:** It took a lifetime, three trips to the moon and the downfall of communism to make it happen...
>
> **[OUT]:** EN-GB,EN-US
>
> **[IN]:** ...as an artist, there is no shortage of colour in my life.
>
> **[OUT]:** EN-GB
>
> **[IN]:** ...the annual pop culture event bringing colorful cosplayers, entertainment aficionados and comic book lovers together under one roof...
>
> **[OUT]:** EN-US

The English varieties contains 3 classes, `EN-US`, `EN-GB` or both `EN-GB,EN-US`. The Portuguese and Spanish varieties also have 3 classes. Respectively, `PT-BR`, `PT-PT` and `PT-BR,PT-PT` for Portuguese from Brazil, Portugal or both and `ES-AR`, `ES-ES` and `ES-AR,ES-ES` for Spanish from Argentina, Spain or both.

For the French varieties, the single label classes comprises the Belgium, Canada, Switzerland and France, viz. `FR-BE`, `FR-CA`, `FR-CH` and `FR-FR`. And the combinations of multi-labels may come from either of the labels, e.g. `FR-CA,FR-CH,FR-FR` to represent texts that could be in classified as either Canadian, Swiss and French varieties. Also, we note that the input text from the French varieties subtask contains masked named-entities represented by the `$NE$` tokens.

[1] The general idea is to generate language labels as how a language model will generate the next token/word in natural text (Li et al., 2018b,c).

[2] https://sites.google.com/view/vardial-2022/shared-tasks

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 230–234
June 20, 2024 ©2024 Association for Computational Linguistics

2 TL;DR (Experimental Setup)

We use the Mistral instruct model with 7 billion parameters (Mistral-7B) (Jiang et al., 2023) for all our experiments .[3]

Off the shelf, we did not *post-train*, i.e. continue training the language model generation with raw monolingual texts, nor *fine-tune* the model with language identification datasets.

Without using the training data, we only selected one example per class from each language family from the development dataset provided by the DSL-ML shared task organizers. These examples were used as one-shot prompt and prepended to texts in the test sets.

For example, given an example from each class in the English development set from Section 1 and a input text from the test set:

> **[IN]:** Conducting an amateur orchestra and performing with it as a soloist are parts of the learning process for young professionals.

We process the above to put them in the format that the Mistral-7B model expects, e.g.

```
<s>[INST] It took a lifetime... [/INST]
EN-GB,EN-US</s>
<s>[INST] ...as an artist... [/INST]
EN-GB</s>
<s>[INST] ...the annual pop culture... [/INST]
EN-US</s>
[INST] Conducting an amateur orchestra... [/INST]
```

And we expected the model to generate `EN-US`, `EN-GB` or `EN-GB,EN-US` as a continuation to the examples and input sentence we entered. We will refer to this as *one-shot prompting* for the rest of the paper.[4] We repeated the *one-shot prompting* approach for the French, Portuguese and Spanish test sets (Zampieri et al., 2024, 2023; Găman et al., 2023; Bernier-colborne et al., 2023).

2.1 One-shot Prompting with Instructions

Additionally, for the English variety test set (Tan et al., 2014a), we experimented with a instruction prompt where we prepend the following instructions before the examples and the test instance, aka. *instructed one-shot prompting*.

```
Label the following text as (i) EN-US if it's
in United States English or (ii) EN-GB if it's
in United Kingdom English or (iii) EN-US,EN-GB
if it can be both in United States or
United Kingdom. <s>[INST] ...[/INST]... </s>...
[INST] Conducting an amateur orchestra... [/INST]
```

3 Results

Lang	Train	Dev	Test
EN	75.1	74.8	74.5
ES	21.2	20.6	21.3
PT	20.0	20.6	18.5
FR	-	15.6	12.9

Table 1: Weighted Averaged F1 Score of One-shot Prompting

Table 1 presents the weighted F1 scores of the one-shot prompting without instructions. In addition to the test set scores, we report the performance of the results of classifying the training (*Train*) and development (*Dev*) of the one-shot prompting approach.

We note that these numbers for the test set F1 scores differ from the ones reported in the official shared task findings papers (Chifu et al., 2024) since we didn't do any special label processing to compute partial matches for multi-class true labels before computing the weighted F1-score with `sklearn`.[5]

Split	One-shot	Prompt-shot
Train	75.1	69.9
Dev	74.8	68.7
Test	74.5	74.8

Table 2: Results of English Variety Classification between One-shot Prompting without (`One-shot`) vs with Instructions (`Prompt-shot`)

Table 2 reports the results of the English variety classification with and without the pre-example instruction prompt as described in Section 2.1. The one-shot prompts with instructions consistently performs worse on the training and development sets as compared to the one-shot prompting without instructions. However, one-shot prompting performs almost equally on F1-scores on the test sets with or without instructions.

[3] https://huggingface.co/mistralai/Mistral-7B-v0.1

[4] We acknowledge that the terminology of "*-shot" has not been defined formally in previous literature, e.g. https://datascience.stackexchange.com/q/120637/122. In this case, we refer to *one-shot* as giving the model one example per class as context before requiring it to infer the label given the test instance.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html

4 Related Works

As language coverage of identification systems increases (Jauhiainen et al., 2019; Agarwal et al., 2023; Burchell et al., 2023), language identification between similar languages, dialects and national varieties remains an active and challenging task in NLP (Tiedemann and Ljubešić, 2012; Gaman et al., 2020; Bouamor et al., 2019; Chakravarthi et al., 2021; Aepli et al., 2022b, 2023).

Early studies on language varieties classification created annotations through proxy signals such using the top-level domain of the text source's website as the locale label (Tan et al., 2014b). However, datasets with locale labels created through proxy signals are often unreliable since there might be no linguistics marker that distinguish one language variety to another language variety (Zampieri et al., 2014; Ács et al., 2015; Goutte et al., 2016).

Zampieri et al. (2023) and Bernier-colborne et al. (2023) redefined the language variety identification task as a multi-label task instead of assigning only a single language variety to each text.

5 Conclusion

By prompting the Mistral-7B model, which was not particularly known to be trained on language identification, we were able to make it classify language varieties to some extent. However, like many large language models, it is largely English-centric and we observed that the English variety classification performance far exceeds the French, Portuguese or Spanish varieties classification task. While a language model '*open source*' its model parameters, the lack of transparency in what goes into training the model makes its usage a grey-box probing exercise.[6]

References

Judit Ács, László Grad-Gyenge, and Thiago Bruno Rodrigues de Rezende Oliveira. 2015. A two-level classifier for discriminating similar languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 73–77, Hissar, Bulgaria. Association for Computational Linguistics.

Noëmi Aepli, Antonios Anastasopoulos, Adrian Chifu, William Domingues, Fahim Faisal, Mihaela Găman, Radu Tudor Ionescu, and Yves Scherrer. 2022a. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, Gyeongju, Republic of Korea. International Committee on Computational Linguistics (ICCL).

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022b. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Milind Agarwal, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. LIMIT: Language identification, misidentification, and translation using hierarchical models in 350+ languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14496–14519, Singapore. Association for Computational Linguistics.

Gabriel Bernier-colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Gaman Mihaela, 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*, pages 1–11, Kiyv, Ukraine. Association for Computational Linguistics.

[6] https://gist.github.com/alvations/af68bd50d4e59d4e74f3632d9ce44e7c (Tan, 2023)

Adrian Chifu, Goran Glavaš, Radu Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City. 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).

Nat Gillin. 2022. Is encoder-decoder transformer the shiny hammer? In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 80–85, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating similar languages: Evaluations and explorations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1800–1807, Portorož, Slovenia. European Language Resources Association (ELRA).

Mihaela Găman, Adrian-Gabriel Chifu, William Domingues, and Radu Tudor Ionescu. 2023. Frecdo: A large corpus for french cross-domain dialect identification. *Procedia Computer Science*, 225:366–373. 27th International Conference on Knowledge Based and Intelligent Information and Engineering Sytems (KES 2023).

Yonatan Hadar and Erez Shmueli. 2021. Categorizing items with short and noisy descriptions using ensembled transferred embeddings. *arXiv preprint arXiv:2110.11431*.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: a survey. *J. Artif. Int. Res.*, 65(1):675–682.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Maggie Yundi Li, Stanley Kok, and Liling Tan. 2018a. Don't classify, translate: Multi-level e-commerce product categorization via machine translation.

Maggie Yundi Li, Stanley Kok, and Liling Tan. 2018b. Don't classify, translate: Multi-level e-commerce product categorization via machine translation. *CoRR*, abs/1812.05774.

Maggie Yundi Li, Liling Tan, Stanley Kok, and Ewa Szymanska. 2018c. Unconstrained product categorization with sequence-to-sequence models. In *Proceedings of the Workshop on eCommerce (co-located with SIGIR)*, pages 1–6.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Liling Tan. 2023. Transparent, opaque and translucent open source llms. *alvations.com*.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014a. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15, Reykjavik, Iceland.

Liling Tan, Marcos Zampieri, Nikola Ljubešić, and Jörg Tiedemann. 2014b. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 6–10. Workshop on Building and Using Comparable Corpora (BUCC) ; Conference date: 27-05-2014 Through 27-05-2014.

Khin Yee Mon Thant and Khin Thandar Nwet. 2020. Comparison of supervised machine learning models for categorizing e-commerce product titles in myanmar text. In *2020 International Conference on Advanced Information Technologies (ICAIT)*, pages 194–199. IEEE.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India. The COLING 2012 Organizing Committee.

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023. Large language models are zero-shot text classifiers.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2023. Language variety identification with true labels.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2024. Language variety identification with true labels. In *Proceedings of LREC-COLING*.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

A Appendix

All generations from the Mistral models used to produce the results from Table 1 and 2 can be found on https://huggingface.co/collections/alvations/jelly-shots-662f2661e4a1f7302a85488a