Evaluating Training Data Construction Strategies for Token-Level Language Identification

Emma Bednaříková and Pavel Rychlý

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{536251, pary}@mail.muni.cz

Abstract. This paper contributes to research on developing token-level language identification tool. The tool is designed to recognize Czech and languages frequently spoken in Czechia, such as Slovak and Ukrainian, while also covering additional languages. In this study, multiple datasets are created using three distinct strategies. The datasets are further used to fine-tune a pre-trained language model, and the resulting models are evaluated on datasets containing code-switching.

Keywords: Language identification; Code-switching; Data augmentation.

1 Introduction

Code-switching is a widely practised linguistic phenomenon. The term refers to the alteration between codes (languages or dialects) within one utterance or written text [5]. Historically, code-switching has mostly been observed in multilingual areas during informal speech. However, due to the rise of social media, it is becoming increasingly prevalent in written form online [6].

Certain applications, such as preprocessing text for speech synthesis, filtering data for machine translation or labelling data in specific corpora, require the language of each token in the input to be detected in order to identify potential code-switching. Determining the languages of code-switched text requires a deeper analysis than is supported by commonly utilized language identification tools, which provide only a single language label for the entire input.

The model utilized in this work is a language identification tool that operates on the token level and can therefore be used to identify code-switching. The tool is based on a pre-trained model and fine-tuned on a synthetically created dataset. In this work, we aim to determine which strategy of creating the training dataset yields the best results when evaluated on real examples of code-switching.

2 Model description

This work compares multiple models developed using the same training strategy and differing only in the datasets used for training. The models use

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2025, pp. 45–53, 2025.

mBERT [3] as their underlying pre-trained model[1]. Alternatives to these models can be developed by utilizing a different pre-trained model. However, since we aim to compare the impact of the strategy used to create the training dataset rather than the capacity of the underlying pre-trained model, all models compared in this work are solely fine-tuned versions of mBERT.

Each entry of the training datasets consists of a list of tokens representing the input and a list of language labels representing the desired predictions. Figure 1 presents an example of an entry from one of the training datasets.

```
{
    'tokens': [
        'Stále', 'se', 'pokoušejí', 'určit', ',', 'jak',
        'velký', 'was', 'the', 'impact', 'and', 'jaký',
        'bude', 'mít', 'vliv', 'na', 'Zemi', '.'
    ],
    'labels': [
        'ces', 'ces', 'ces', 'ces', 'ces', 'ces',
        'ces', 'eng', 'eng', 'eng', 'eng', 'ces',
        'ces', 'ces', 'ces', 'ces', 'ces', 'ces'
    ]
}
```

Fig. 1: An example of an entry from one of the training datasets.

The tokens were processed by the mBERT's tokenizer and since the tokenizer splits the tokens in the entry into smaller units, the labels required to be realigned to match the tokenizer's outputs. For the training, the batch size was set to 16, the number of epochs was 3, and the learning rate was 2e-5. The training was conducted on one GPU Tesla T4.

3 Training datasets

To build the training datasets, data from the FLORES-200 dataset[1] were collected[2]. Despite, this dataset being usually utilized as a machine translation benchmark, we opted for it, as it represents a reliable multilingual source with shared domain across the languages. The FLORES-200 dataset contains parallel sentences from more than 200 languages. From these we selected only sentences from a subset of languages, namely: Arabic, Czech, Danish, German, English, French, Haitian Creole, Italian, Japanese, Dutch, Polish, Portuguese,

[1] https://huggingface.co/google-bert/bert-base-multilingual-cased
[2] `https://github.com/facebookresearch/flores/tree/main/flores200`

Russian, Slovak, Spanish, Swedish, and Ukrainian. These raw data were further augmented using different strategies to compose the training datasets.

In total, 13 datasets were developed that can be categorized into three distinct strategies. In the first strategy, the raw data were first split into words using the `word_tokenize` method from the `nltk.tokenize` [3] library. For Japanese text a different tokenizer[4] had to be utilized as `word_tokenize` does not support Japanese. Secondly, we took a part of the tokenized sentences and used the data to create snippets of different lengths (the lengths varied depending on the specific dataset type). These snippets were further inserted at random positions in to the remaining of the sentences. In this paper, this strategy will be reffered to as *random* strategy.

In total, 9 datasets were created using this strategy. The datasets differ in the maximal number of snippets inserted in one sentence and the maximal length a snippet can have. The last variable indicates whether the snippets of length 1 are being preferred over snippets of other lengths. If set to *yes*, 30 % of the snippets are of length 1. Otherwise, the lengths of the snippets are distributed equally between 1 and the maximum range. The inclusion of this variable was motivated by the observation that, in many cases of code-switching, speakers switch from language A to language B only to use a single word before returning to language A. By analyzing two datasets containing code-switching, we observed that 30% of the islands (snippets of text written in a different language than the remainder of the sentence), were a single word. Therefore, in some datasets, 30% of the inserted snippets are of length 1, and the remaining lengths are distributed equally within the remaining 70 %. Datasets composed using this strategy are presented in Table 1.

Table 1: Datasets created by inserting snippets into sentences randomly.

dataset name	**max snippets count**	**max snippet length**	**weighted**
random-S1-L12-w	1	12	yes
random-S1-L12-e	1	12	no
random-S1-L06-w	1	6	yes
random-S1-L06-e	1	6	no
random-S1-L03-w	1	3	yes
random-S1-L03-e	1	3	no
random-S3-L06-w	3	6	yes
random-S3-L06-e	3	6	no
random-S3-L03-e	3	3	no

[3] `https://www.nltk.org/api/nltk.tokenize.word_tokenize.html`
[4] `https://pypi.org/project/fugashi/`

In the second strategy, the input sentences were tokenized as in the first strategy. However, the snippets inserted into a sentence were obtained from translations of the same sentence. Thus, each snippet shares a context with the sentence into which it is being inserted. This strategy was utilized to create two datasets. This strategy will be in this paper reffered to as *parallel* strategy.

The last strategy aimed to mimic the most the real nature of code-switching. Consequently, the modification of the sentence was achieved not by inserting random snippets into the original sentence, rather by selecting a piece of the original sentence and translating it. The part for the translation was selected randomly, and the length of the translated part varied within a specific range. For the translation, we used `googletrans`[5] library, as it is a freely available translation tool that supports all languages we selected to include in the model. We utilized this strategy to create two training datasets; their overview, along with the overview of datasets created using the *parallel* strategy, is presented in Table 2. This strategy will be reffered to as *translated* strategy.

Table 2: Datasets created by inserting snippets from parallel sentences into the original sentence.

dataset name	**max snippets count**	**max snippet length**	**weighted**
parallel-s	1	6	yes
parallel-l	1	10	no
translated-s	1	6	yes
translated-l	1	10	no

The following are examples of tokens in three different entries, where each entry represents a different strategy of creating the training dataset:

– *random*

`However` `,` `they` `are` `not` `thought` `to` `have` `caused` `any`
`damage` `to` `the` `co` `nejdříve` `abyste` `získali` `shuttle` `.`

– *parallel*

`Stále` `se` `pokoušejí` `určit` `,` `jak` `velký` `byl` `náraz`
`determine` `just` `how` `large` `a` `jaký` `bude` `mít` `vliv` `na`
`Zemi` `.`

[5] `https://pypi.org/project/googletrans/`

– *translated*

Die	ersten	Fälle	choroba	tohto	Saison	wurden	Ende
Juli	gemeldet	.					

In the first example, the tokens `co`, `nejdříve`, `abyste`, `získali` are from a Czech sentence and were inserted into an English sentence without any syntactical or semantic context. In the second example, the tokens `determine`, `just`, `how`, `a` were inserted from an English equivalent of the surrounding Czech sentence. In the last example, the tokens `choroba`, `tohto` are Slovak translations of tokens on the fourth and fifth position of the surrounding German sentence.

4 Evaluation

The models created by fine-tuning mBERT on datasets described in Section 3 were evaluated on two publicly available datasets. The first dataset used for evaluation of the models is the Denglisch dataset[4], and the second is the Bangor Miami corpus[2].

Denglisch dataset The Denglisch dataset is a collection of real-life examples of German-English code-switching. The data in this dataset were obtained from the social network Reddit. The authors of this dataset provided token-level annotations and their annotation scheme contains 7 categories: English, German, Mix, Shared English, Shared German, Shared Other, and Other. The evaluation was conducted only on tokens labeled as English or German, and the models' predictions for tokens that fell into the remaining categories were ignored. The following is an example of tokens in an entry in the Denglisch dataset.

Absolut	well	played	kann	man	nicht	anders	sagen	.

Bangor Miami corpus The Bangor Miami corpus is a collection of token-level annotated transcripts of audio recordings. Since the speakers in the audio recordings are fluent in both English and Spanish, code-switching between these two languages is frequently present. The tokens in this corpus were divided into three categories: English, Spanish, and English/Spanish (for tokens whose language cannot be derived). We evaluated the models only on tokens that were labeled as English or Spanish, tokens from the last category were ignored. The following are tokens from an entry in the Denglisch dataset.

ay	dios	mío	how	embarrassing	!

The tokenizer of the fine-tuned model splits most of the tokens received in the input into subtokens. Therefore, a strategy that merges the results from the subtokens provided by the model into the original tokens in the evaluation datasets was needed. For the purpose of this evaluation, we added the models' predictions for each language across the subtokens, and the language with the highest sum of logits was selected as the language of the given token. This strategy is illustrated in Figure 2 to enhance comprehension. The values representing the model's predictions are only symbolic.

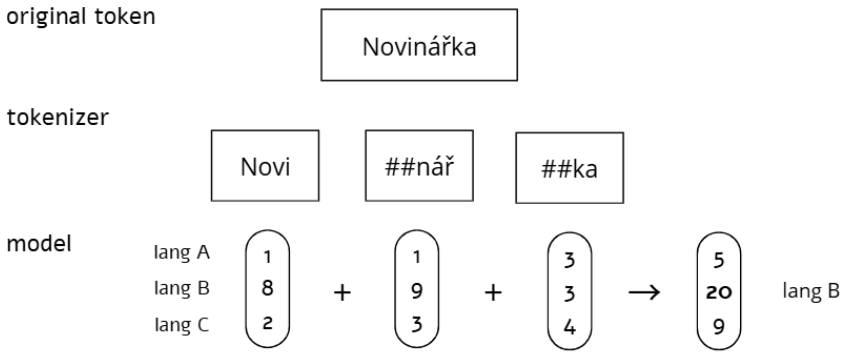

Fig. 2: An illustration of the strategy used to merge the outputs provided by the models from subtokens into tokens.

All variants of the model were evaluated on the Denglisch dataset and the Bangor Miami corpus. The accuracies of the model variants obtained from the evaluation are presented in Table 3 and Figure 3.
The model variant that achieved the highest score on the Denglisch dataset was the model trained on the dataset *random-S1-L12-e*, while the highest performing model variant on the Bangor Miami corpus was the variant trained on the dataset *random-S1-L03-w*. The model variant with the highest average of accuracies achieved on the two utilized datasets was the variant trained on the *parallel-l* dataset.

The differences between the model results were minimal, and none of the models consistently outperformed the others across both datasets. Therefore, it is not possible to clearly determine the best approach to constructing the training dataset. An interesting observation is that the models trained on datasets constructed using the *parallel* and *translated* strategies performed better on variants with longer snippets than on those with shorter ones. Another interesting observation is that the model trained on the *translated-s* dataset, which theoretically best mimics real code-switching, performed the worst on the Bangor Miami corpus and the second worst on the Denglisch dataset, resulting in the lowest average accuracy among all models.

Table 3: The accuracies of different model variants achieved from the evaluation on Denglisch dataset and Bangor Miami corpus.

model variant	Denglisch dataset	Bangor Miami corpus	Average
random-S1-L12-w	0.9672	0.9503	0.9588
random-S1-L12-e	**0.9707**	0.9403	0.9555
random-S1-L06-w	0.9667	0.9501	0.9584
random-S1-L06-e	0.9678	0.9467	0.9573
random-S1-L03-w	0.9616	**0.9530**	0.9573
random-S1-L03-e	0.9646	0.9414	0.9530
random-S3-L06-w	0.9673	0.9488	0.9581
random-S3-L06-e	0.9681	0.9384	0.9533
random-S3-L03-e	0.9654	0.9378	0.9516
parallel-s	0.9648	0.9433	0.9541
parallel-l	0.9691	0.9503	**0.9597**
translated-s	0.9643	0.9358	0.9501
translated-l	0.9704	0.9470	0.9587

The inconsistency in results between the model variants on the Denglisch dataset and the Bangor Miami corpus can be attributed to differences between the two datasets. Besides the fact that one contains English and German while the other contains English and Spanish, the datasets also differ in the style of language they contain. Although both comprise informal language, the Denglisch dataset consists of informal yet written posts, whereas the Bangor Miami corpus consists of transcriptions of audio recordings and therefore contains many filler words, such as *so*, *hm*, *no*, *um*, whose language is without context challenging to recognize.

5 Conclusion

In this study, we presented three distinct approaches for generating training datasets. We then fine-tuned a pre-trained model on each of these datasets. We then evaluated the resulting models on two datasets containing code-switching. Our aim was to determine the most effective strategy for creating the training dataset. However, the results did not reveal a clear winner, as all the models produced similar outcomes.

Acknowledgments. This work has been partly supported by the Ministry of Education, Youth and Sports of the Czech Republic within the LINDAT-CLARIAH-CZ project LM2023062.

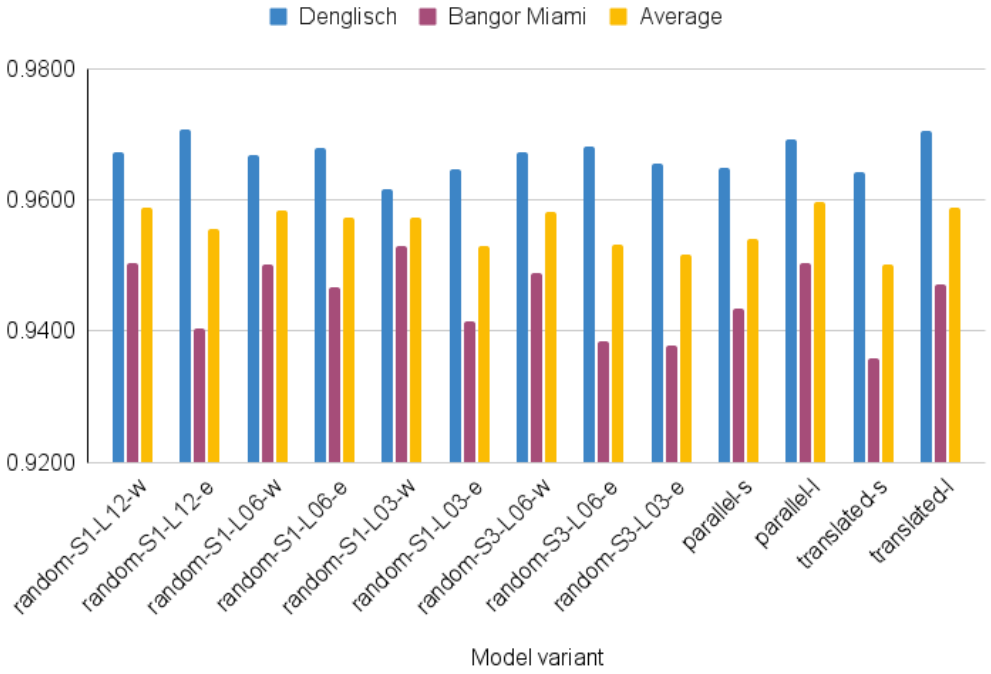

Fig. 3: The accuracies of different model variants achieved from the evaluation on Denglisch dataset and Bangor Miami corpus and their average values.

References

1. Costa-jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G.M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K.R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N.F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., Wang, J., NLLB Team: Scaling neural machine translation to 200 languages **630**(8018), 841–846. https://doi.org/10.1038/s41586-024-07335-x, https://doi.org/10.1038/s41586-024-07335-x
2. Deuchar, M.: BilingBank spanish-english bangor miami corpus. https://doi.org/10.21415/T5J01D, https://biling.talkbank.org/access/Bangor/Mi\protect\discretionary{\char\hyphenchar\font}{}{}ami.html
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) 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). pp. 4171–4186. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423/
4. Osmelak, D., Wintner, S.: The denglisch corpus of German-English code-switching. In: Beinborn, L., Goswami, K., Muradoğlu, S., Sorokin, A., Kumar, R., Shcherbakov, A., Ponti, E.M., Cotterell, R., Vylomova, E. (eds.) Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP.

pp. 42–51. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023). https://doi.org/10.18653/v1/2023.sigtyp-1.5, https://aclanthology.org/2023.sigtyp-1.5/
5. Spolsky, B.: Sociolinguistics, vol. 1. Oxford university press
6. Zhang, K., Yang, L.: Analysis of code-switching phenomena in the evolution of social media language—take the mixed use of chinese and english by gen z as an example. Journal of Humanities, Arts and Social Science **9**(3) (2025)