

LyS A Coruña at GUA-SPA@IberLEF2023: Multi-Task Learning with Large Language Model Encoders for Guarani-Spanish Code Switching Analysis

Alberto Muñoz-Ortiz¹, David Vilares¹

¹Universidade da Coruña, CITIC, Departamento de Ciencias de la Computación y Tecnologías de la Información, Campus de Elviña s/n, 15071, A Coruña, Spain

Abstract

This paper introduces the LyS A Coruña proposal for the Guarani-Spanish Code Switching Analysis task at IberLEF2023. The shared task proposes to analyze Guarani-Spanish code-switched texts, focusing on language identification, named entity recognition (NER), and a novel classification task for Spanish spans in a code-switched Guarani-Spanish context. We propose three multi-task learning systems that have common encoders based on two language models and different decoders in a multi-task learning setup. The encoders use the contextual embeddings by: (i) a large language model (LLM) pretrained on bidirectional machine translation on 200 languages (including Spanish and Guarani) from the No Language Left Behind project, and (ii) a BERT-based model pretrained in Spanish and finetuned in around 800k Guarani tokens. The decoders are: (i) a softmax output layer for Task 1, and (ii) conditional random fields (CRF) output layers for Tasks 2 and 3. According to official results, we ranked third in the three tasks.

Keywords

Multi-Task Learning, Guarani, Spanish, Code-switching, Language identification, Named Entity Recognition, Code Classification,

1. Introduction

Indigenous and European languages have coexisted in South America for centuries, causing lexical and cultural interchanges to different degrees. In Paraguay, where both Guarani and Spanish are recognized as official languages and the population is mostly bilingual, code-switching has emerged in this bilingual environment. This bilingual environment has favored the apparition of code-switching. Code-switching is a linguistic phenomenon where people alternate between two or more languages in the same conversation, which is common in places where more than one language interact [1].

Code-switching is not a new topic in natural language processing (NLP), including language identification [2], named entity recognition (NER) [3], sentiment analysis [4] or machine translation [5]. The goal of the Guarani-Spanish Code Switching Analysis shared task [6] at IberLEF 2023 [7] is to automatically analyze code-switching between Guarani and Spanish


IberLEF 2023, September 2023, Jaén, Spain

✉ alberto.munoz.ortiz@udc.es (A. Muñoz-Ortiz); david.vilares@udc.es (D. Vilares)

🌐 <https://amunozo.github.io/> (A. Muñoz-Ortiz); <https://grupolys.org/~david.vilares/> (D. Vilares)

🆔 0000-0001-9608-2730 (A. Muñoz-Ortiz); 0000-0002-1295-3840 (D. Vilares)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

in texts from news and social media. Particularly, they propose three tasks: (i) language identification, (ii) named entity recognition and (iii) Spanish code classification.

The main challenge of this task is the disparity of resources of the two languages involved. While Spanish is one of the most rich-resourced languages in the world, resources in Guarani are very scarce due to the lack of labeled and unlabeled data [8].

Our approach We follow a multi-task learning approach using MaChAmp [9]. We trained three different models using the contextual embeddings of two models available at [https://huggingface.co/: facebook/nllb-200-distilled-600M](https://huggingface.co/facebook/nllb-200-distilled-600M), a large language model (LLM) pretrained on bidirectional translation across 200 languages (including Guarani and Spanish), and `mmaguero/beto-gn-base-cased`, a BERT-based LLM pretrained in Spanish data and finetuned in approximately 800k Guarani tokens. We train three independent models, one single-task on Task 1 and two multi-task on the three tasks. We used the `nllb` to train a single-task model on Task 1 that uses a softmax output layer on top of the encoder, and a multi-task model that is trained on the three tasks, adding a softmax output layer for Task 1, and two independent masked conditional random fields (CRF) decoders for Tasks 2 and 3 that force the labels to follow the BIO schema. We used the `beto-gn` model to train a multi-task learning model with the same three decoders as the multi-task `nllb` model. A post-processing heuristic is added to the predicted file to make sure the tags combinations are assigned correctly. According to official results, we ranked third in the three tasks.

2. Overview of the shared task

The shared task aims to analyze code-switched texts in Guarani and Spanish through three token-level tasks: language identification in code-switched data, named entity classification, and Spanish code classification. The specific tasks are:

- **Task 1: Language identification in code-switched data.** Given a sentence, this task consists in identifying the language of the token or a named entity. The possible labels are `gn` (Guarani word), `es` (Spanish word), `ne` (part of a named entity), `mix` (mixture between Guarani and Spanish), `foreign` (words from neither Guarani and Spanish) and `other`.
- **Task 2: Named entity classification.** The task goal is to identify the tokens that span named entities in the text and classify them into three different categories. It follows the BIO schema (first token of an entity is marked with B- plus the label, and the following starting with I-, and lastly the tokens that are not part of an entity are marked with an O). The different labels of this tasks are: `ne-b-loc` and `ne-i-loc` for location, `ne-b-org` and `ne-i-org` for organization, and `ne-b-per` and `ne-i-per` for person.
- **Task 3: Spanish code classification.** The task goal is label the spans of text in Spanish depending on its form: `es-b-cc` and `es-i-cc` represent change in code, i.e. Spanish tokens that keep all its original characteristics of Spanish, while `es-b-u1` and `es-i-u1` represent not-adapted loans, i.e. Spanish words that have been partial adapted to Guarani syntax.

Metrics For Task 1, the metrics used are accuracy, weighted precision, weighted recall, and the main metric is weighted F1. For Tasks 2 and 3, the metrics are precision, recall and F1, unlabeled and labeled, being the latter the main one. The criterion used for considering a named entity valid is exact match.

Data We only used the dataset provided by the competition to train our models. The dataset consists in sentences annotated at token level with the previously introduced labels. There are 19 003 training tokens in 1 140 training sentences, 2 989 development tokens in 180 development sentences and 2 857 test tokens in 180 test sentences.

3. Our models

We rely on MaChAmp [9], a toolkit that allows to use contextualized embeddings from HuggingFace’s [10] LLMs in multi-task learning setups. Multi-task learning (MTL) [11, 12] consists in learning more than one task at once with the goal of achieving a better generalization. The most common approach and the one we use is called hard parameter sharing [13]. It consists of sharing the hidden layers between all tasks while using a task-specific decoder for each one. MTL helps model generalize, as representations must be general and flexible enough to learn different tasks at the same time, reducing the risk of overfitting.

Our models uses a pre-trained large language model as a shared encoder for all tasks, and independent decoders for each of the tasks. For Task 1, we use a softmax output layer on the output of the encoder’s embeddings, while for Tasks 2 and 3 we use a masked conditional random field [14] that force the output labels to follow the BIO schema. During training, the layers of the encoder are updated for all of the tasks, while the decoders’ layers are only trained with the loss of the corresponding task.

The final results of the shared task are chosen independently of the systems, that means that if one submission had a better result in Task 1 and another in Task 2, these two will appear in the final result. What we are reporting here are the independent results of each model in each task. In our case, each category was topped by a different system, all of them using the same approach described, but with differences in encoders and decoders. We trained three different models that were tested independently, one trained only on Task 1 and two trained on the three tasks:

1. `single-nllb`. The model that performed the best in Task 1 uses `facebook/nllb-200-distilled-600M` [15] as the encoder, a large language model with 600 million parameters, pretrained in machine translation across 200 languages (including Spanish and Guarani), for a total of 40 602 total translation directions, and a softmax output layer as a decoder for Task 1. This model has not been trained on Tasks 2 and 3.
2. `mtl-nllb`. With the same encoder as before, the model that obtained the best results in Task 3 is trained on the three tasks instead of just one, using a softmax output layer for Task 1 and CRFs output layers for Tasks 2 and 3.
3. `mtl-beto-gn`. The model that obtained the highest results in Task 2 uses the same decoders as `mtl-nllb`, but uses the contextual embeddings generated by

mmaguero/beto-gb-base-cased [16]. This LLM is a BERT-based model pretrained in Spanish and finetuned on approximately 800K Guarani tokens.

3.1. Pre-processing

In the official Shared Task description, Tasks 2 and 3 are represented as a single label together with *ne* and *es* labels respectively. However, as we predict each label with an independent decoder, we represent them independently: tagsets are {*gn*, *es*, *ne*, *mix*, *foreign*, *other*} for Task 1, {*B-loc*, *I-loc*, *B-org*, *I-org*, *B-per*, *I-per*, *O*} for Task 2 and {*B-cc*, *I-cc*, *B-ul*, *I-ul*, *O*} for Task 3, so for example an original label *es-b-cc* would be represented as (*es*, *O*, *B-cc*) in our data.

3.2. Post-processing

As the three tasks are predicted independently, we introduced a post-processing step with a heuristic in order to ensure the correctness of the output labels. As Task 2 can only go together with named entities in Task 1 and Task 3 with Spanish tokens, we had to modify the output labels that did not match these requirements with a simple heuristic:

1. If the token is tagged for Task 2 and the label for Task 1 is not *ne*, we changed label from Task 1 to *ne*. The same happens for Task 3, if the label for Task 1 is not *es*, we change label from Task 1 to *es*. For example, (*foreign*, *O*, *B-cc*) would be change to (*es*, *O*, *B-cc*)
2. If the token is both labeled for Task 2 and 3, which should not happen, we decide which one remains depending on Task 1: if label 1 is *ne*, we delete label 3, and if label 2 is *es*, we delete label 2. For example, (*es*, *B-org*, *B-cc*) would be changed to (*es*, *O*, *B-cc*).
3. If label 1 is neither *ne* nor *es* and the token is labeled for Task 2 and 3, we delete randomly Task 2 or Task 3, and change Task 1 to the correspondent label for the remaining task. For example, (*foreign*, *B-org*, *B-cc*) could be transformed into (*es*, *O*, *B-cc*) or (*ne*, *B-org*, *O*).

For an illustrative example, see Table 1.

Input token	Output	Post-processed output
Horacio	(<i>es</i> , <i>B-per</i> , <i>O</i>)	(<i>ne</i> , <i>B-per</i> , <i>O</i>)
Melanio	(<i>gn</i> , <i>I-per</i> , <i>O</i>)	(<i>ne</i> , <i>I-per</i> , <i>O</i>)

Table 1

Post-processing for the invalid tokens Horacio Melanio.

3.3. Training details

Our three models have been trained using the same hyperparameters during 100 epochs on a Tesla A100 GPU. Hyperparameters are listed in Table 2.

Hyperparameter	Value
"batch_size"	16
"diverse"	"false"
"max_tokens"	8192
"sampling_smoothing"	1
"shuffle"	"true"
"sort_by_size"	"true"
"default_dec_dataset_embeds_dim"	12
"encoder_dropout"	0.3
"max_input_length"	128
"update_weights_encoder"	"true"
"keep_top_n"	1
"cut_frac"	0.3
"decay_factor"	0.1
"discriminative_fine_tuning"	"true"
"gradual_unfreezing"	"true"
"lr"	2e-05
"num_epochs"	100

Table 2
Training hyperparameters used for our models.

User	Task 1 - weighted F1	Task 2 - Labeled F1	Task 3 - Labeled F1
pughrob	0.9381 (1)	0.7028 (1)	0.3836 (1)
tsjauhia	0.9139 (2)	-	-
amunozo (us)	0.8500 (3)	0.4153 (3)	0.1939 (3)
<i>baseline</i>	<i>0.7325 (4)</i>	<i>0.4946 (2)</i>	<i>0.2195 (2)</i>
pakapro	0.452 (5)	-	-

Table 3
Final results for the Guarani-Spanish Code Switching Analysis at IberLEF 2023.

4. Results

The overall results of the competition are shown in Table 3. Detailed results of the runs for our three models are shown in Tables 4, 5 and 6.

	Accuracy	W. Precision	W. Recall	W. F1	M. Precision	M. Recall	M. F1
Task 1	0.8530	0.8502	0.8530	0.8500	0.6699	0.6056	0.6294

Table 4
Detailed results for the system `single-n11b`. W. stands for weighted and M. stands for macro.

As we can see in Table 3, our model is only able to surpass the baseline for Task 1, while being a bit under it for Tasks 2 and 3. We believe our models are unable to generalize well for these tasks, as they were able to learn the training test easily but they could not obtain good results on the development and test sets. There are two main reasons that can make this happen:

	Accuracy	W. Precision	W. Recall	W. F1	M. Precision	M. Recall	M. F1
Task 1	0.8288	0.8205	0.8288	0.8241	0.5667	0.5510	0.5566
	L. Precision	L. Recall	L. F1	U. Precision	U. Recall	U. F1	
Task 2	0.4021	0.3744	0.3878	0.4709	0.4384	0.4541	
Task 3	0.2500	0.1584	0.1939	0.3360	0.2129	0.2606	

Table 5

Detailed results for the system mt1-n11b. In Task 1, W. stands for weighted and M. stands for macro. In Tasks 2 and 3, L. stands for labeled and U. stands for unlabeled.

	Accuracy	W. Precision	W. Recall	W. F1	M. Precision	M. Recall	M. F1
Task 1	0.8295	0.8236	0.8295	0.8226	0.588	0.5429	0.5604
	L. Precision	L. Recall	L. F1	U. Precision	U. Recall	U. F1	
Task 2	0.4663	0.3744	0.4153	0.5337	0.4286	0.4754	
Task 3	0.2824	0.1188	0.1674	0.3411	0.1436	0.2021	

Table 6

Detailed results for the system mt1-beto-gn. In Task 1, W. stands for weighted and M. stands for macro. In Tasks 2 and 3, L. stands for labeled and U. stands for unlabeled.

- The small quantity of data makes it challenging to obtain a representation that generalizes well over the task. The modest amount of data for Guarani in the pretrained representations does not allow the model to learn better representations that would have helped on Tasks 2 and 3 and avoided overfitting.
- Second, the selected models were the only models on HuggingFace that had data on Guarani and Spanish. However, their big number of parameters could lead to more overfitting. We tried to control this by using different rates of dropout in both the encoder and the decoders, but we could not obtain a satisfactory result. We also tried some smaller models with data in Guarani, but the lack of Spanish data in the contextual embeddings harmed the final results.

5. Conclusion

Our paper describes LyS A Coruña contribution to the GUA-SPA Code Switching Analysis at IberLEF2023. We participated in the three tasks, applying a multi-task learning setup with hard-sharing and an independent decoder for each task. As encoders, we used a linear decoder for Task 1 and a CRF both for Tasks 1 and 2 to make sure the predicted labels follow the BIO schema. According to official results, we ranked third in the three categories, not being able to surpass the baselines for Tasks 2 and 3, due to the lack of generalization of our proposed model.

Acknowledgments

This paper has received funding from grant SCANNER-UDC (PID2020-113230RB-C21) funded by MCIN/AEI/10.13039/501100011033, grant FPI 2021 (PID2020-113230RB-C21) funded by MCIN/AEI/10.13039/501100011033, the European Research Council (ERC), which has supported

this research under the European Union’s Horizon Europe research and innovation programme (SALSA, grant agreement No 101100615), Xunta de Galicia (ED431C 2020/11), and Centro de Investigación de Galicia “CITIC”, funded by Xunta de Galicia and the European Union (ERDF - Galicia 2014-2020 Program), by grant ED431G 2019/01.

References

- [1] A. K. Joshi, Processing of sentences with intra-sentential code-switching, in: *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*, 1982. URL: <https://aclanthology.org/C82-1023>.
- [2] T. Solorio, E. Blair, S. Maharjan, S. Bethard, M. Diab, M. Ghoneim, A. Hawwari, F. AlGhamdi, J. Hirschberg, A. Chang, P. Fung, Overview for the first shared task on language identification in code-switched data, in: *Proceedings of the First Workshop on Computational Approaches to Code Switching*, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 62–72. URL: <https://aclanthology.org/W14-3907>. doi:10.3115/v1/W14-3907.
- [3] G. Aguilar, F. AlGhamdi, V. Soto, M. Diab, J. Hirschberg, T. Solorio, Named entity recognition on code-switched data: Overview of the calcs 2018 shared task, *arXiv preprint arXiv:1906.04138* (2019).
- [4] P. Patwa, G. Aguilar, S. Kar, S. Pandey, S. Pykl, B. Gambäck, T. Chakraborty, T. Solorio, A. Das, Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets, *arXiv preprint arXiv:2008.04277* (2020).
- [5] S. Chen, G. Aguilar, A. Srinivasan, M. Diab, T. Solorio, Calcs 2021 shared task: Machine translation for code-switched data, *arXiv preprint arXiv:2202.09625* (2022).
- [6] L. Chiruzzo, M. Agüero-Torales, G. Giménez-Lugo, A. Alvarez, Y. Rodríguez, S. Góngora, T. Solorio, Overview of GUA-SPA at IberLEF 2023: Guarani-Spanish Code-Switching Analysis, *Procesamiento del Lenguaje Natural 71* (2023).
- [7] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*, co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org, 2023.
- [8] P. Joshi, S. Santy, A. Budhiraja, K. Bali, M. Choudhury, The state and fate of linguistic diversity and inclusion in the NLP world, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020, pp. 6282–6293. URL: <https://aclanthology.org/2020.acl-main.560>. doi:10.18653/v1/2020.acl-main.560.
- [9] R. van der Goot, A. Üstün, A. Ramponi, I. Sharaf, B. Plank, Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP, in: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Online, 2021, pp. 176–197. URL: <https://aclanthology.org/2021.eacl-demos.22>. doi:10.18653/v1/2021.eacl-demos.22.
- [10] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf,

- M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: <https://aclanthology.org/2020.emnlp-demos.6>. doi:10.18653/v1/2020.emnlp-demos.6.
- [11] R. Caruana, Multitask learning, *Machine learning* 28 (1997) 41–75. URL: <https://link.springer.com/article/10.1023/A:1007379606734>.
- [12] S. Ruder, An overview of multi-task learning in deep neural networks, arXiv preprint arXiv:1706.05098 (2017). URL: <https://arxiv.org/abs/1706.05098>.
- [13] R. Caruana, Multitask learning: A knowledge-based source of inductive bias¹, in: Proceedings of the Tenth International Conference on Machine Learning, Citeseer, 1993, pp. 41–48.
- [14] J. Lafferty, A. McCallum, F. C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001).
- [15] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, J. Wang, No language left behind: Scaling human-centered machine translation, 2022. arXiv:2207.04672.
- [16] L.-H. A. G. Agüero-Torales, Marvin Matías, D. Vilares, Multidimensional affective analysis for low-resource languages: A use case with guarani-spanish code-switching language, *Cognitive Computation* (2023).