Multi-Varietal Text Classification for Quechuan Languages

Claire Benet Post
University of Colorado Boulder
benet.post@colorado.edu

1 Introduction

Quechua, an indigenous language family widespread across the Andean region, exhibits a significant linguistic diversity with over 40 varieties, spoken by millions across South America (Luykx et al., 2016; Hornberger and King, 1998; Grimes, 1985; Adelaar, 2020). This diversity, while enriching, presents substantial challenges for computational linguistics, particularly in tasks like text classification, machine translation, and interlinear glossing (IGT) (Buys and Botha, 2016; Himoro and Lora, 2022; Wiemerslage et al., 2022), due to the wide-ranging orthographic and dialectal variations (Hornberger and Limerick, 2019; Limerick, 2018; Sarasola et al., 2020).

This project endeavors to address a critical gap in computational linguistics for Quechua: the development of a multi-varietal text classifier. Existing methodologies, such as the binary classification system developed by Medina (Medina, 2013), only distinguish texts as either Cuzco-Quechua or non-Cuzco-Quechua. In contrast, the goal of this project is to build a classifier capable of distinguishing between multiple Quechua varieties. This endeavor is pivotal for enhancing the granularity of dialectical recognition within the language family and lays the groundwork for a more nuanced computational model for multi-varietal Quechua morphological parsing, which is the focus of my upcoming preliminary paper for my PhD.

Given Quechua's status as a low-resource language (Cardenas et al., 2018), this work is crucial, requiring innovative use of linguistic analyses and computational methods to identify and classify the nuanced linguistic features unique to Quechua dialects. This approach not only seeks to enhance dialectal recognition within the Quechua language family but also contributes to the broader goal of creating tailored computational tools for linguistically diverse but digitally under served languages.

My planned methodology includes trying to leverage QuBERT, a pre-trained model tailored for Southern Quechua (Zevallos et al., 2022a). This project will employ a combination of semi-supervised and unsupervised methods for a refined approach to text classification. By fine-tuning QuBERT on a dataset annotated with dialectal information, the model will extend its capabilities to multiple Quechua varieties. Semi-supervised strategies, such as self-training and co-training, will enhance the dataset with high-confidence predictions, while unsupervised clustering uncovers dialectal patterns, enriching the model's dialectal sensitivity.

However, discerning unique linguistic traits across Quechua varieties presents challenges, notably due to the scarcity of digital resources. This project will employ unsupervised learning, data augmentation, and explore cross-linguistic parallels to overcome these hurdles. Contingency measures, such as refining feature extraction or experimenting with alternative classifiers, are prepared should initial strategies falter. Support from existing corpora (Melgarejo et al., 2022; Cardenas et al., 2018) and resources like QuBERT (Zevallos et al., 2022a), along with efforts to amass a broader corpus, will bolster this endeavor.

In conclusion, this project not only aims to expand the horizons of Quechua computational resources but also aspires to contribute to NLP for low-resource languages at large. Achieving successful multi-varietal classification for Quechua could aid linguistic analysis and enhance processing capabilities for downstream applications.

2 Related Work

In this section related work to my project will be reviewed in addition to providing details on similarities and differences between my approach and others. I plan to build upon previous work within the space of text classification, but the key challenge is the low-resource setting of Quechua. Thus,

I detail how I will utilize other methodologies in this section.

The paper (Medina, 2013) represents a significant milestone in the realm of computational linguistics for Quechua. Published in 2013, this work pioneers the development of an automatic text classifier for Quechua dialects, marking the first attempt at automated dialect classification within the language family. Employing attributes such as words, lemmas, bigrams, and trigrams, the study focuses on achieving a binary distinction between Cusco Quechua and non-Cusco Quechua dialects. Despite this binary focus, the paper lays the groundwork for future endeavors by highlighting the potential for extending the classification framework to encompass more nuanced distinctions between dialects.

The project encountered several challenges reflective of the low-resource nature of Quechua, notably the scarcity of electronically available written documents suitable for classification purposes. Furthermore, the inherent dialectal variability within Quechua posed a significant challenge, complicating the classification task. The paper underscores the importance of preprocessing techniques, such as stop word lists and lemmatization, in improving classifier performance, emphasizing the necessity of a curated corpus for effective automated classification in low-resource languages like Quechua.

Looking ahead, the paper outlines avenues for future research, including the expansion of the corpus to cover more dialects and the exploration of sophisticated machine learning techniques tailored to capture the linguistic characteristics of Quechua more effectively. Additionally, it highlights practical considerations such as the standardization challenge in Peru and the need for web applications to facilitate automatic classification of Quechua texts, particularly those from the Cusco dialect.

My project expands upon Medina's paper by attempting the future work noted on multi-dialectical Quechua classification. I plan on using the research and insights found within the paper within my own work, while updating my methodology to incorporate modern NLP techniques.

Next, the paper (Zevallos et al., 2022b) addresses the challenge of resource scarcity for Quechua by presenting a large corpus tailored for deep learning of the southern dialect. This work introduces QuBERT, a pre-trained transformer model based on RoBERTa, specifically trained on this curated corpus using the best-performing normalization and tokenization techniques. QuBERT represents the largest linguistic model for Quechua to date, comprising nearly 450,000 segments. Additionally, the paper introduces a normalization technique based on finite-state transducers (FSTs) to enhance the quality and consistency of the corpus. Various tokenization techniques, including byte-pair encoding (BPE), BPE-Guided, and Prefix-Root-Postfix-Encoding (PRPE), are applied to the corpus, with each technique made available for download to cater to the diverse preferences and needs of researchers. Furthermore, the paper evaluates QuBERT's performance on named-entity recognition (NER) and part-of-speech (POS) tagging tasks, achieving high accuracy.

My project plans to leverage QuBERT by fine-tuning it on a dataset annotated with dialectal information, extending its capabilities to multiple Quechua varieties. This approach, combined with semi-supervised and unsupervised methods, aims to enhance text classification for Quechua, enriching the model's dialectal sensitivity.

In the paper (Alammary, 2022), this systematic review examines various BERT models applied to Arabic text classification, emphasizing their effectiveness and performance compared to other machine learning models. Although focusing on Arabic, the study provides insights into the general applicability of BERT models for text classification tasks in low-resource languages. My project draws inspiration from the methodologies and insights outlined in this review, adapting them to the context of Quechua multi-varietal text classification.

Finally, the (Shnarch et al., 2022) paper proposes a method to enhance the performance of pre-trained models, like BERT, for text classification tasks, especially in scenarios with limited labeled data. While my project does not specifically focus on cold start scenarios, the concept of leveraging unsupervised clustering for enhancing classification performance aligns with the semi-supervised and unsupervised methods I aim to incorporate for Quechua multi-varietal text classification.

References

Willem FH Adelaar. 2020. Morphology in quechuan languages. In *Oxford Research Encyclopedia of Linguistics*.

Ali Saleh Alammary. 2022. Bert models for arabic text

classification: a systematic review. *Applied Sciences*, 12(11):5720.

Jan Buys and Jan A Botha. 2016. Cross-lingual morphological tagging for low-resource languages. *arXiv preprint arXiv:1606.04279*.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua. *ISI-NLP*, 2:21.

Joseph E Grimes. 1985. The interpretation of relationships among quechua dialects. *Oceanic Linguistics Special Publication*, pages 271–284.

Marcelo Yuji Himoro and Antonio Pareja Lora. 2022. Preliminary results on the evaluation of computational tools for the analysis of quechua and aymara. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5450–5459.

Nancy H Hornberger and Kendall A King. 1998. Authenticity and unification in quechua language planning. *Language Culture and Curriculum*, 11(3):390–410.

Nancy H Hornberger and Nicholas Limerick. 2019. Teachers, textbooks, and orthographic choices in quechua: Bilingual intercultural education in peru and ecuador. In *Perspectives on Indigenous writing and literacies*, pages 141–164. Brill.

Nicholas Limerick. 2018. Kichwa or quichua? competing alphabets, political histories, and complicated reading in indigenous languages. *Comparative Education Review*, 62(1):103–124.

Aurolyn Luykx, Fernando García Rivera, and Félix Julca Guerrero. 2016. Communicative strategies across quechua languages. *International Journal of the Sociology of language*, 2016(240):159–191.

Rosemary Jiménez Medina. 2013. *Clasificación Por Dialecto De Documentos Escritos En Quechua*. Ph.D. thesis, Universidad Nacional De San Antonio Abad.

Nelsi Melgarejo, Rodolfo Zevallos, Hector Gomez, and John E. Ortega. 2022. WordNet-QU: Development of a lexical database for Quechua varieties. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4429–4433, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kepa Sarasola, Iñaki Alegria, and Olatz Perez-De-Viñaspre. 2020. Language technology for language communities: An overview based on our experience (2020). *Language and Technology in Wales: Volume I*, page 10.

Eyal Shnarch, Ariel Gera, Alon Halfon, Lena Dankin, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2022. Cluster & tune: Boost cold start performance in text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7639–7653, Dublin, Ireland. Association for Computational Linguistics.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann. 2022. Morphological processing of low-resource languages: Where we are and what's next. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 988–1007, Dublin, Ireland. Association for Computational Linguistics.

Rodolfo Zevallos, Luis Camacho, and Nelsi Melgarejo. 2022a. Huqariq: A multilingual speech corpus of native languages of Peru forSpeech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5029–5034, Marseille, France. European Language Resources Association.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Nuria Bel, Cesar Toshio, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. 2022b. Introducing qubert: A large monolingual corpus and bert model for southern quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13.