Introducing QuBERT: A Large Monolingual Corpus and BERT Model for Southern Quechua

Rodolfo Zevallos[◇] **John E. Ortega**[§] **William Chen**[▽] **Richard Castro**[Ω] **Nuria Bel**[◇]
Cesar Yoshikawa[ψ] **Renzo Ventura**[ψ] **Hilario Aradiel**[ψ] **Nelsi Melgarejo**[α]
[◇]Universitat Pompeu Fabra [§]Universidade de Santiago de Compostela (CITIUS)
[▽]University of Central Florida [Ω]Universidad Nacional de San Antonio Abad
[ψ]Universidad Nacional del Callao [α]Pontificia Universidad Católica del Perú
{rodolfojoel.zevallos, nuria.bel}@upf.edu, john.ortega@usc.gal, wchen6255@knights.ucf.edu,
rcastro@hinant.in, {ctyoshikawaa, rventuras, haradielc}@unac.edu.pe, nelsi.melgarejo@pucp.edu.pe

Abstract

The lack of resources for languages in the Americas has proven to be a problem for the creation of digital systems such as machine translation, search engines, chat bots, and more. The scarceness of digital resources for a language causes a higher impact on populations where the language is spoken by millions of people. We introduce the first official large combined corpus for deep learning of an indigenous South American low-resource language spoken by millions called *Quechua*. Specifically, our curated corpus is created from text gathered from the southern region of Peru where a dialect of Quechua is spoken that has not traditionally been used for digital systems as a target dialect in the past. In order to make our work repeatable by others, we also offer a public, pre-trained, BERT model called *QuBERT* which is the largest linguistic model ever trained for any Quechua type, not just the southern region dialect. We furthermore test our corpus and its corresponding BERT model on two major tasks: (1) named-entity recognition (NER) and (2) part-of-speech (POS) tagging by using state-of-the-art techniques where we achieve results comparable to other work on higher-resource languages. In this article, we describe the methodology, challenges, and results from the creation of QuBERT which is on on par with other state-of-the-art multilingual models for natural language processing achieving between 71 and 74% F1 score on NER and 84–87% on POS tasks.

1 Introduction

With the availability of online digital resources for computation and data storage, the capability for executing natural language processing (NLP) tasks such as named-entity recognition (NER), part-of-speech (POS) tagging, and machine translation (MT) on low-resource languages, languages with few digital resources available, has increased. The processing power and data available for experimentation are unsurpassed in history and research (Edwards, 2021) has shown that in the current decade we are on track to overcome previous methods, such as Moore's law (Schaller, 1997), for predicting computing time of experiments. This finding is better observed on high-resources languages like English and French where the amount of data that exists is more than enough to take advantage of the latest computing architectures. Unfortunately, for other low-resource languages like Quechua, an indigenous language spoken by millions in Peru, South America, it is more difficult to create statistically significant NLP models due to the amount of data needed (typically on the order of millions of sentences). Therefore, it is critical to create public-facing mechanisms for low-resource languages like Quechua to help provide research collaboration which will improve the quality for low-resource language NLP systems. We aim to improve the digital resources available for Quechua by curating a large monolingual corpus for southern Quechua, a dialect of Quechua spoken in the southern region of Peru not commonly found in most literature.

The initiative we present in this article can be considered a major contribution and advancement as means to improve the quality of NLP tasks for the Quechua language. We outline the multiple innovations and contributions provided below.

1. A considerably large, curated, monolingual corpus of southern Quechua consisting of nearly 450K segments.

2. A normalization technique applied to the corpus based on finite-state transducers (FSTs) (Rios, 2015; Rios and Göhring, 2016; Ortega et al., 2020a).

Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing, pages 214 - 226

3. Several tokenization techniques applied to the corpus, each made available for download, including byte-pair encoding (BPE) (Sennrich et al., 2015), BPE-Guided (Ortega et al., 2020a), and Prefix-Root-Postfix-Encoding (PRPE) (Chen and Fazio, 2021; Zuters et al., 2018).

4. A pre-trained transformer model based on RoBERTa (Liu et al., 2019) called *QuBERT* that uses the corpus along with the best performing normalization and tokenization techniques from items 2 and 3 above.

5. A comparison of the performance of the techniques introduced in items 2 and 3 above on a NER classification task.

6. A comparison of the performance of the techniques introduced in items 2 and 3 above on a POS classification task.

In order to cover our innovations and contributions, we highlight the details in several sections. First, in Section 2, we describe the latest work on Quechua and other techniques related to low-resource NLP tasks such as the ones we introduce on NER and POS. Next in Section 3, we provide more background on the Quechua language by covering morphological, phonological, and other important grammatical details. Then, we describe how we curated our corpus in Section 4. In Section 5, we provide details on the parameters and configuration for our models and tokenization techniques which leads way to the experimental evaluation and results from the NER and POS tasks in Section 6. Finally, we wrap up with a few proposed lines of future work and a conclusion in Section 7.

2 Related work

In this section we present several works that can be considered state-of-the-art at this time for Quechua. Since we are introducing several new contributions, we briefly cover the most recent work and how it related to each contribution mentioned.

First, concerning the introduction of the corpus, we discuss work where corpora have been introduced for public use. Like many low-resource NLP projects, one of the several corpora that is often used is the Opus[1] (Tiedemann, 2012) corpus. It contains text similar to ours in southern Quechua (Quechua II, see more details on Quechua variants in Section 3); however, it contains biblical text only. Other work (Ortega et al., 2020a) introduced the JW300 corpus (Agić and Vulić, 2019); their corpus was for one domain also. The corpus we present contains entries from several diverse sources while at the same time including Opus and the JW300. Ortega et. al (Ortega et al., 2020a) also presented a magazine selection known as *Hinantin* which contained 250 non-biblical Quechua—Spanish sentences found on-line[2]. While the Hinantin magazine was a more diverse domain than other Quechua corpora previously introduced, our corpus is the largest and most diverse compiled currently available.

Our second contribution consists of a normalization technique used in previous work (Rios, 2015; Rios and Göhring, 2016; Ortega et al., 2020a). The work presented in this article uses the same normalization technique (described further in Section 5) but, to our knowledge, this is the first time that the normalization technique has been used on a corpus of this size for southern Quechua.

Thirdly, for Quechua, there has not been a tokenization comparison similar to the one presented here. There are two works (Chen and Fazio, 2021; Ortega et al., 2020a) that present approaches called *BPE-Guided* and *PRPE* separately but their work did not compare on such a varied corpus for named-entity recognition or part-of-speech tasks, both of their works for the machine translation task only.

The fourth, fifth, and sixth contributions are all related to the first-time presentation of a deep learning transformer model for Quechua that is used for NER and POS classification tasks. One of the works that presented deep learning approaches for Quechua is a shared task (Mager et al., 2021a) from the first workshop on NLP for indigenous languages of the America (Mager et al., 2021b). Another work called *indt5* (Nagoudi et al., 2021) used an encoder-decoder model transformers based on T5 (Raffel et al., 2020). Both models were mainly used for translation and the data did not contain nearly as much Quechua–Spanish text as ours. (Ortega et al., 2020a) applied a deep learning approach where quality was low due to the use of the Opus corpus for training and Hinantin for test – their deep learning approach was for machine translation also. Other work (Zheng et al., 2021; Liu et al., 2020) has presented large corpora with trans-

[1] http://opus.nlpl.eu

[2] http://hinant.in

former architectures but did not include Quechua as one of the low-resource languages. The one work that can be considered closest to ours in size and technique is the work by Wongso et. al (Wongso et al., 2021), they pre-trained mono-lingual models on GPT-2 (Radford et al., 2019), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019). Like our work, they used a monolingual corpus which consisted of a variety of text and evaluated the models on a sentiment classification task for Sudanese. The main difference between their work and our work is that our tasks are slightly different and are based on Quechua. In order to better understand why NLP tasks for Quechua can be more complex than for other languages, we present more details in the next section on the language.

3 Quechua language

Quechua is an indigenous language native to several regions in South America, mainly Peru, Ecuador, and Bolivia, and is spoken by nearly 8 million people. It is known (Pinnis et al., 2017; Kann, 2019; Karakanta et al., 2018) to be a highly inflective language based on its suffixes which agglutinate. Due to its morphology, Quechua has been found to be similar to other languages like Finnish (Ortega et al., 2021, 2020b; Ortega and Pillaipakkamnatt, 2018).

Linguistically, Quechua can be considered a unique and even complex language due to the highly polysynthetic nature and phonology. Slight changes in morphemes (small sub-word units) can modify a word's meaning drastically. Since Quechua is the South American language with the highest amount of native speakers and those speakers tend to introduce diverse accentuated tones on different words depending on the locality, one can assume that the combination of morphological and tonal rules that cause inflection can make tasks like the ones presented in this article (NER and POS) difficult due to the high likelihood of non-common meanings for sub-words and letters. For example, by adding an accent to the letter 'o' in Quechua, words become plural.

Quechua synthesis, or the *synthetic index* (Greenberg, 1963) – the average number of morphemes per word, is about two times larger than English. English typically has around 1.5 morphemes per word and Quechua has about 3 morphemes per word. This high morphological complexity has been described in detail in the past (Muysken, 1988); few have been able to overcome the challenges that low-resource languages like Quechua present for digital processing. Quechua's phonology uses three vowels for the most part: *a*, *i*, and *u*. Consonants, on the other hand, are numerous and depending on the region where it is spoken, Quechua can have up to 14 constants (Ortega et al., 2020a). Generally speaking, lexemes are monosyllabic or bi-syllabic having two vowels (VV) or two consonants (CC) that do not concur in the same syllable. From a phonological perspective, the scheme of any Quechua root is: (C)V(C)-CV(C) (Cerrón-Palomino, 1994).

The region where Quechua is spoken can be considered important. Alfredo Torero (Torero, 1964) reported that there are two main divisions of the language (Quechua I and Quechua II). Quechua II is mostly spoken in regions such as Ayacucho, Peru and is considered a "southern" language. There are several more dialects spoken and others (Adelaar, 2004) report several divisions for Quechua II; but, in this article we focus specifically on the southern version at a high-level.

A lot of the Quechua morphology has been documented in previous works (Rios et al., 2008; Rios, 2015; Muysken, 1988; Monson et al., 2006; Torero, 1964); however, there is not a clear consensus to resolve all morphology issues that may arise. In order to statistically determine which branch of morphemes a verb phrases falls under can be difficult with Quechua since there are so few resources. A short example sentence of how complex morpheme determination can be is depicted in Table 1. In some cases, there are hundreds of options to choose from when choosing which suffix to use for a given Quechua word.

4 Corpus details

4.1 Monolingual

We consider the introduction of our monolingual corpus on southern Quechua the largest corpus of its kind to date. Table 4 gives a precise overview of all of the corpora that we have combined in October 2021 in order to present our corpus publicly online[3]. We have created the corpus from several sources. The majority of corpora combined to create the final corpus is a compilation of 50 monolingual corpora from different sources on the web including OSCAR (Suárez et al., 2019), JW300 (Agić and

[3] https://huggingface.co/datasets/Llamacha/monolingual-quechua-iic

Test sentence: **Chanta**pis **Biblia**nejta **qotucha**kuynejta **ima yanapalla**wanchejtaj

Stemmed Morpheme	Potential Suffixes
Chanta	–pis –s
Biblia	–niq –ta
qutachu	–ku –y –niq –ta
ima	
yanapa	–lla –wa –nchik –ta
yanapalla	–wa –nchik –ta

Table 1: The sub-segment suffix choices of a short sentence for a Quechua sentence. (Ortega et al., 2020a)

Vulić, 2019), and CC-100 (Conneau et al., 2020; Wenzek et al., 2020). To our knowledge, these corpora have not yet been introduced as one southern Quechua corpus to the wider research community. Additionally, our corpus contains other corpora mentioned below (see Table 4 for a complete list) that are not easily found on-line.

The introduction of our corpus is part of a larger project called *Llamacha*[4] focused on helping under-resourced communities . In Llamacha, the authors have begun to use the corpus directly as a form of creating software tools able to help teachers in regions of southern Peru where Quechua II is spoken. Llamacha tools cover several use cases such as government documents, children's internet tools, and more. This demand constitutes the main reason we distribute this corpus for public use – it is our hope that others from the research community will get involved to help develop more tools that can use our corpus.

With such a high demand for diverse performance, we compiled our corpus to cover the domains mentioned and more. Our compilation spans across several domains including religion, economics, health, social, political, justice and culture. We consulted several sources such as books and stories from Andean narratives and the Peruvian Ministry of Education[5] to collect data. Table 4 illustrates the entire data set which consists of 4,408,953 tokens and 384,184 sentences, including what are known as "Chanka" and "Collao" variants, variants specific to the Quechua II branch. In effect, we have created a corpus that is nearly ten times larger than most widely used Quechua corpus (Rios, 2015) until now which has eight combined corpora, 47,547 tokens, and 3,614 sentences.

4.2 Named-entity recognition and part-of-speech

Both the NER and POS corpora were created using the corpus introduced and are made publicly available online[6]. There are slight differences, nonetheless, between the amount of examples used that we note in this section.

In order to create the NER and POS corpora a team of ten annotators were selected. The annotators were university students and 7 of 10 of them were native Quechua speakers. Nonetheless, they were all students of what is known as a "Intercultural Bilingual Education" in Peru where students are taught coursework in both Quechua and Spanish. Annotation was performed using Label-Studio[7] to annotate sentences for NER and POS.

The NER corpus was built using 5,450 sentences using the CoNLL2003 (Sang and De Meulder, 2003) format. Work was reviewed to ensure that annotations were standardized and using an BIO format annotating only the following tags: Person (PER), Location (LOC) and Organization (ORG). The POS corpus was built using 4,229 sentences and annotated identical to previous work on POS Rios (2015) for Quechua. Additionally, as a way of having a more precise tagging strategy, we used official dictionaries of "Chanka" and "Collao" Quechua from the Peruvian Ministry of Education to identify POS tag correctness.

5 Experimental settings

5.1 Tokenization

Our tokenization strategy is to include the state-of-the-art techniques currently being used for Quechua, regardless if it is Quechua I or II (Torero, 1964). We do this as a mechanism to show that

[4] https://llamacha.pe
[5] http://www.minedu.gob.pe/
[6] https://github.com/Llamacha/QuBERT
[7] https://labelstud.io

Text	Ismael Montes Hatun Yachay Wasi Yachachiqkunap
BPE	Ismael Montes H@@atun Yachay Wasi Yachachiqkuna@@p
PRPE	Ismael Monte@@s Hatun Ya@@chay Wasi Yach@@achiq@@kuna@@p
BPE-Guided	Is@@m@@a@@el Mon@@t@@es Hatun Yachay Wasi Yach@@achiq@@kunap

Table 2: The use of four word-tokenization techniques for Quechua.

the corpus presented in Section 4 can be used to achieve high performance (around 80–90% accuracy) for tasks similar to high-resource languages as a recent survey (Li et al., 2020) has shown.

We use the latest tokenization techniques which focus on sub-word segmentation. (Haddow et al., 2021; Chen and Fazio, 2021; Ortega et al., 2020a; Sennrich et al., 2015) Byte-pair encoding (BPE) (Sennrich et al., 2015) can be considered one of the most widely-used approaches and a fundamental technique that has served as a baseline for previous research (Ortega et al., 2021, 2020a,b) on Quechua. The BPE approach is considered the de-facto standard tokenization algorithm for agglutinative languages (Chimalamarri and Sitaram, 2021). BPE represents text at the character-level and then merges the most frequent pairs iteratively until a pre-determined number of merge operations have been reached. Our BPE tokenizer was trained on the entire collective corpus from Section 4 with a vocabulary size of 52,000.

Alternatively, we additionally experiment with a popular extension of the BPE technique called *BPE-Guided* (Ortega et al., 2020a), used for increasing performance on Quechua machine translation. BPE-Guided is similar to the BPE approach in that it iteratively "discovers" sub-word segmentation by jointly learning a vocabulary and character-level segmentation. The extension offered by BPE-Guided is that it introduces Quechua knowledge in a *a-priori* manner by using the BPE algorithm for excluding common suffixes found on Wikimedia[8] before learning a vocabulary or segmentation. In our experiments, we use the list of Quechua suffixes introduced previously (Ortega et al., 2020a).

Another tokenization technique that has been shown to perform better than BPE and BPE-Guided on Quechua texts (Chen and Fazio, 2021) is known as the Prefix-Root-Postfix-Encoding (PRPE) (Zuters et al., 2018) technique. The PRPE algorithm separates words into three main divisions: (1) a prefix, (2) a root, and (3) a postfix. It completes this separation by first learning a sub-word vocabulary through detecting potential prefixes and post-fixes based on a heuristic. It then aligns the prefixes and post-fixes into sub-strings of a word to find potential roots. Once the roots have been located, the text is segmented into sub-words according to their statistical probability. Table 2 shows an example southern Quechua sentence tokenized by the three approaches mentioned.

Lastly, all text with exception of one experiment (Text and BPE in Table 3) is normalized with the Quechua toolkit (Rios, 2015) that uses finite-state transducers (Mohri, 1997) to determine if words belong to the same category and can be merged into one. Rios (2015)[Section 2.5] describe their normalization methodology which contains four models that are based on morphology, the "normalization" technique used in our experiments follows their work which includes all four models.

5.2 Model Architecture

We call our model **QuBERT** because it is a transformer model based on BERT (Devlin et al., 2019). More specifically, our model has been trained using the RoBERTa (Devlin et al., 2018) enhancement to BERT which can be considered higher-performing for NER and POS tasks (Li et al., 2020). An example of the model architecture is shown in Figure 1 which shows how our model produces NER classifications given a Quechua sentence.

Our model has been first pre-trained with southern Quechua text on 384,184 sentences. Then, we fine tuned the model with 4,360 sentences for the NER task and 3,383 sentences for the POS task. For the training process, we used 6 hidden layers. Each layer was 768 dimensions, giving us a total of 84 million parameters. For optimization, we used the Adam optimizer with hyper-parameter values of β_1=0.9 and β_2 = 0.99 along with a learning rate of 2.7e-06. Lastly, we incorporated a weight

[8] https://en.wiktionary.org/wiki/Category:Quechua_suffixes

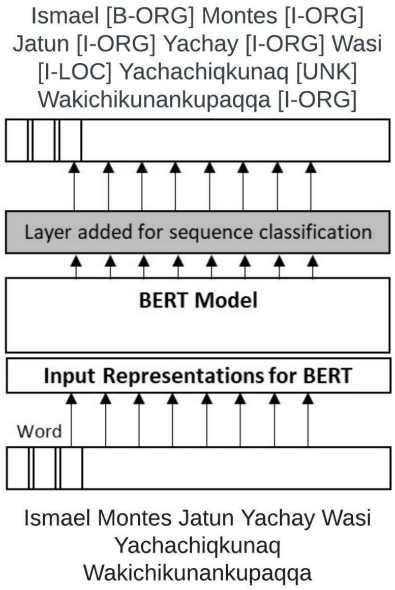

Figure 1: Model architecture based on Bert (Devlin et al., 2019).

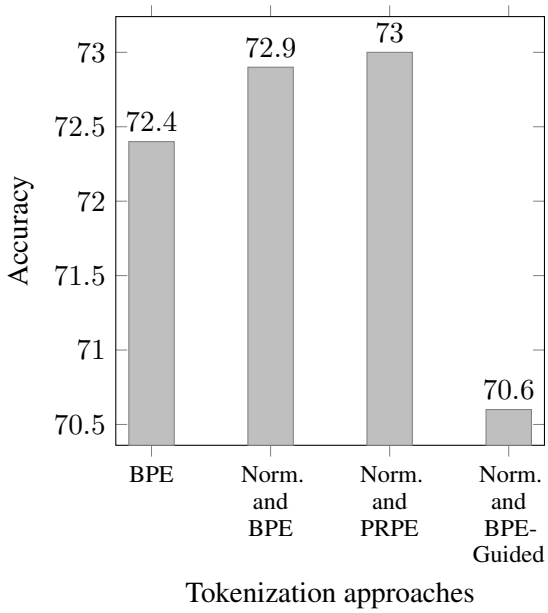

Figure 2: An accuracy comparison of tokenization techniques on southern Quechua (Quechua II) using a RoBERTa (Liu et al., 2019) model for named-entity recognition (NER).

decay factor of 0.1 to prevent overfitting. The pre-training was for two epochs and a batch size of 64 with 12k iterations, before being fine-tuned on the downstream task for 10 epochs and a batch size of 32. Initial development was done on a Google Colab[9] notebook, while models used for final testing were pre-trained and fine-tuned on a single 16GB NVIDIA Tesla V100 GPU.

6 Results

The results presented in this section show how well **QuBERT** performs on two main tasks: NER and POS. We feel that the contributions presented in Section 1 are sufficient to warrant wider use of our work; however, it is our intention to show that the corpus, model, and experiments could provide easy access for future work. We cover each task (NER and POS) as separate sections below in order to provide better insight into how the model performs in different scenarios, specifically for the different tokenization and normalization (called "norm." in Table 3) techniques mentioned in Section 5. Nonetheless, we provide precision, recall, and F1 scores in Table 3 for both tasks as an aggregate to get an overall sense of how well our base model performs on both tasks.

6.1 Named Entity Recognition

Figure 2 illustrates the accuracy from our model on the NER task. We note that the accuracy scores are somewhat lower than the state-of-the-art for high-resource languages on the NER task (Li et al., 2020). However, our F1 scores seems to be inline with other newly published work on low resources (Bouabdallaoui et al., 2022) (69–70%) for various deep learning models). In future work, we plan on adapting our model to more complex architectures such as those found in SemEval-2022 Task 11 (Malmasi et al., 2022).

To further investigate the findings we report the following findings[10] based on these NER tags: B-LOC, B-ORG, B-PER, I-LOC, I-ORG, I-PER, O. When text was normalized and then tokenized with BPE we noticed that I-ORG and I-PER were the highest amount of true positives (227 and 196 respectively) when compared to other tokenization techniques. However, BPE without normalization performed worse than other techniques on I-PER classification, mainly classifying them as B-LOC. BPE-Guided generally scored similar to BPE on NER with a trend of being slightly lower than BPE. PRPE scored better on I-LOC and I-ORG (306 and 227 respectively) than other techniques and was able to achieve the highest accuracy of all techniques.

From the illustration in Figure 2, we believe that

[9] https://colab.research.google.com

[10] For a complete confusion matrix, please refer to Appendix Table 5.

Tokenization Approach	NER			POS		
	F1	Prec	Recall	F1	Prec	Recall
Text and BPE	0.736	0.749	0.724	0.860	0.859	0.862
Text with norm. and BPE	0.741	0.753	0.729	0.861	0.861	0.862
Text with norm. and PRPE	0.741	0.753	0.730	0.867	0.866	0.868
Text with norm. and BPE-Guided	0.716	0.726	0.707	0.843	0.843	0.843

Table 3: A comparison of tokenization techniques on southern Quechua (Quechua II) using a RoBERTa (Liu et al., 2019) model for classification. Normalization (norm.) is applied using the Quechua toolkit (Rios, 2015). Scores are calculated at the token level and weighted-averaged by class.

the different techniques are closely related but it is clear that the BPE-Guided approach was not as successful for the NER task as it has been in the past for machine translation (Ortega et al., 2020a). We feel that this is probably due to the amount of data introduced in our corpus which did not contain as many matching suffixes as was done in the previous work (Ortega et al., 2020a). Since this is a first-time introduction of a deep learning model for NER in Quechua, we believe that this can serve as a baseline for future work.

6.2 POS tagging

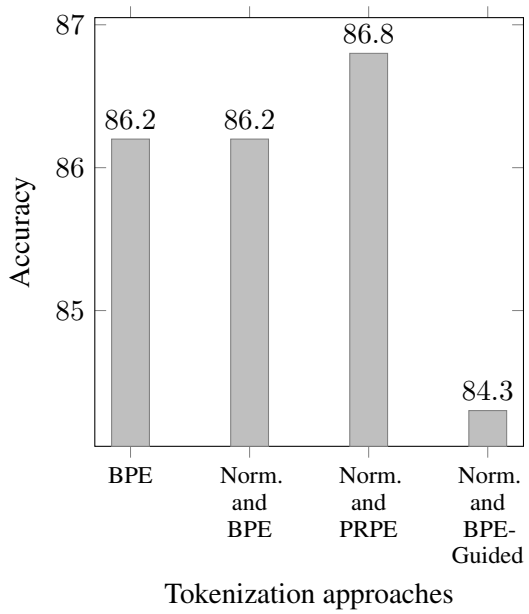

Figure 3: An accuracy comparison of tokenization techniques on southern Quechua (Quechua II) using a RoBERTa (Liu et al., 2019) model for part-of-speech (POS) tagging.

.

The part-of-speech task seems to be more fitted for our model since we are able to achieve accuracy in the high 80% range as shown in Figure 3, similar to other high-resource tasks(Li et al., 2020). We feel that for POS tagging our model is optimal given the current state-of-the-art. Also, our annotations, while completed by a near-native speaker were somewhat easier to complete due to the more rigid classification of vocabulary-based words in Quechua, essentially the annotator could look up words and parts of speech when there was doubt. In the future, as with the NER task, we fill that we can achieve higher quality with professional translators/annotators.

For POS tagging, unlike the NER task, we were able to discern performance from our analysis based on terms that could be found in a dictionary.[11] Adjectives, verbs and adverbs were mostly correct by all tokenization techniques. Particularly, PRPE outperformed other techniques with the correct classification of 262 adjectives when compared to BPE (259) and BPE-Guided (235). PRPE also performed slightly better on POS verb identification than other techniques. BPE-Guided, on the other hand, performed better with determinant detection finding 43 true positives as opposed to 39 by PRPE and BPE.

7 Conclusion and future work

In this article, we have introduced a novel monolingual corpus, curated and compiled for southern Quechua. We have shown that the corpus can be used for downstream tasks such as NER and POS tagging by creating and releasing a deep learning model based on BERT (Devlin et al., 2019) called **QuBERT**. Additionally, we experimented with the state-of-the-art tokenization techniques for preprocessing and normalization in order to achieve results similar to those found on high-resource languages.

[11] For a complete confusion matrix, please refer to Appendix Table 6.

In the future, we would like to experiment with other model architectures for more complex NER tasks such as those presented at SemEval-2022 (Malmasi et al., 2022), of particular interest is the work from Wang et al. (2022). We would like to include more native Quechua speaking annotators in order to improve the data set even more. The introduction of two or more annotators will allows us to introduce models for tasks such as machine translation, question-answering, and topic modeling where the reference data is even more important. We believe that our work can serve as a baseline for future work and invite other researchers to use the contributions presented here for further investigative lines such as the ones we are considering: online tools for native Quechua speakers and human interaction.

Acknowledgments

This work was partially funded by Project PID2019-104512GB-I00 of the Spanish Ministerio de Ciencia, Innovación and Universidades and Agencia Estatal de Investigación.

References

Willem FH Adelaar. 2004. *The languages of the Andes*. Cambridge University Press.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Ibrahim Bouabdallaoui, Fatima Guerouate, Samya Bouhaddour, Chaimae Saadi, and Mohamed Sbihi. 2022. Named entity recognition applied on moroccan tourism corpus. *Procedia Computer Science*, 198:373–378.

Rodolfo Cerrón-Palomino. 1994. Quechua sureño. diccionario unificado. *Biblioteca Básica Perúana, Biblioteca Nacionál del Peru*.

William Chen and Brett Fazio. 2021. Morphologically-guided segmentation for translation of agglutinative low-resource languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 20–31.

Santwana Chimalamarri and Dinkar Sitaram. 2021. Linguistically enhanced word segmentation for better neural machine translation of low resource agglutinative languages. *International Journal of Speech Technology*, pages 1–7.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Edwards. 2021. Moore's law: what comes next? *Communications of the ACM*, 64(2):12–14.

Joseph Harold Greenberg. 1963. Universals of language.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2021. Survey of low-resource machine translation. *arXiv preprint arXiv:2109.00486*.

Katharina Kann. 2019. Acquisition of inflectional morphology in artificial neural networks with prior knowledge. *arXiv preprint arXiv:1910.05456*.

Alina Karakanta, Jon Dehdari, and Josef van Genabith. 2018. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32(1-2):167–189.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios Gonzales, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, et al. 2021a. Findings of

the americasnlp 2021 shared task on open machine translation for indigenous languages of the americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217.

Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann. 2021b. Proceedings of the first workshop on natural language processing for indigenous languages of the americas. Association for Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. Semeval-2022 task 11: Multilingual complex named entity recognition (multiconer). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). Association for Computational Linguistics*.

Mehryar Mohri. 1997. Finite-state transducers in language and speech processing. *Computational linguistics*, 23(2):269–311.

Christian Monson, Ariadna Font Llitjós, Roberto Aranovich, Lori Levin, Ralf Brown, Eric Peterson, Jaime Carbonell, and Alon Lavie. 2006. Building nlp systems for two resource-scarce indigenous languages: mapudungun and quechua. *Strategies for developing machine translation for minority languages*, page 15.

PC Muysken. 1988. Affix order and interpretation: Quechua.

El Moatez Billah Nagoudi, Wei-Rui Chen, Muhammad Abdul-Mageed, and Hasan Cavusoglu. 2021. IndT5: A text-to-text transformer for 10 indigenous languages. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 265–271, Online. Association for Computational Linguistics.

John Ortega and Krishnan Pillaipakkamnatt. 2018. Using morphemes from agglutinative languages like quechua and finnish to aid in low-resource translation. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 1–11.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020a. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

John E Ortega, Richard Alexander Castro-Mamani, and Jaime Rafael Montoya Samame. 2021. Love thy neighbor: Combining two neighboring low-resource languages for translation. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 44–51.

John E Ortega, Richard Alexander Castro Mamani, and Jaime Rafael Montoya Samame. 2020b. Overcoming resistance: The normalization of an amazonian tribal language. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 1–13.

Mārcis Pinnis, Rihards Krišlauks, Daiga Deksne, and Toms Miks. 2017. Neural machine translation for morphologically rich languages with improved subword units and synthetic data. In *International Conference on Text, Speech, and Dialogue*, pages 237–245. Springer.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Annette Rios. 2015. *A basic language technology toolkit for Quechua*. Ph.D. thesis, University of Zurich.

Annette Rios and Anne Göhring. 2016. Machine learning applied to rule-based machine translation. In *Hybrid Approaches to Machine Translation*, pages 111–129. Springer.

Annette Rios, Anne Göhring, and Martin Volk. 2008. A quechua-spanish parallel treebank. *LOT Occasional Series*, 12:53–64.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Robert R Schaller. 1997. Moore's law: past, present and future. *IEEE spectrum*, 34(6):52–59.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Alfredo Torero. 1964. *Los dialectos quechuas*. Univ. Agraria.

Xinyu Wang, Yongliang Shen, Jiong Cai, Tao Wang, Xiaobin Wang, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang, Kewei Tu, et al. 2022. Damo-nlp at semeval-2022 task 11: A knowledge-based system for multilingual named entity recognition. *arXiv preprint arXiv:2203.00545*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Wilson Wongso, Henry Lucky, and Derwin Suhartono. 2021. Pre-trained transformer-based language models for sundanese.

Francis Zheng, Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2021. Low-resource machine translation using cross-lingual language model pretraining. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 234–240, Online. Association for Computational Linguistics.

Jānis Zuters, Gus Strazds, and Kārlis Immers. 2018. Semi-automatic quasi-morphological word segmentation for neural machine translation. In *International Baltic conference on databases and information systems*, pages 289–301. Springer.

A Appendix

The figures below represent several of the individual differences between corpora and their corresponding language in Table 4 and tokenization approaches for NER and POS in Tables 5 and 6 respectively.

Corpus	# Sentences	# Tokens	Dialect	Year	Dominio
jw300_2013	124,038	1,465,494	Chanka	2013	Religion
wikipedia_2021	96,560	1,009,631	Collao	2021	Miscellaneous
cc100-quechua	86,250	1,206,770	Collao	2018	Miscellaneous
jw300_2017	25,585	294,473	Collao	2017	Religion
microsoft	5,018	60,847	Collao	2021	Norma
que_community_2017	21,139	38,570	Collao	2017	Miscellaneous
tribunal_constitucional	1,148	32,974	Chanka	2021	Justice
tierra_vive	4,731	27,768	Collao	2013	Religion
conectamef	433	20,683	Collao	2016	Economy
unesco	937	16,933	Collao	2020	Program
oscar_quz	491	12,717	Collao	2020	Miscellaneous
constitucion_simplified_quz	999	12,217	Collao	1993	Norma
libro_quechua	781	11,476	Chanka	2002	Agreement
handbook_quy	2,297	11,350	Chanka	2019	Education
dw_quz	325	11,079	Collao	2009	Social
yaku_unumanta	283	10,787	Chanka	2013	Norma
uywaymanta	683	9,231	Collao	2015	Education
maria_mamani	987	9,179	Chanka	2011	Education
anta	451	8,839	Collao	2010	Education
Agreement _nacional_2014	356	8,355	Chanka	2014	Agreement
omnilife	336	8,184	Collao	2017	Health
pasado_violencia	373	8,001	Chanka	2008	Social
cosude_2009-2011_qu	536	7,959	Collao	2011	Social
fondo_monetario_internacional	291	7,227	Collao	2010	Economy
peru_suyupi	449	6,420	Chanka	2014	Education
fundacion_quz	440	5,776	Collao	2008	Social
greg_quz	185	5,505	Collao	2010	Narrative
imayna	250	5,425	Chanka	2008	Social
ahk_1968-2008_quz	391	5,186	Collao	2008	Economy
directiva	355	4,988	Chanka	2014	Resolution
achka	256	4,844	Chanka	2015	Education
cartillas	870	4,674	Chanka	2006	Education
lectura-favorita-chanka-2019	781	4,363	Chanka	2019	Education
lectura-favorita-cusco-2019	769	4,351	Collao	2019	Education
amerindia	321	4,280	Chanka	2000	Education
yachay_qipikuna	464	4,174	Collao	2009	Education
reglamento_simplified_quz	287	4,053	Collao	2008	Norma
focus_2008_quz	243	3,797	Collao	2008	Narrative
poder_judicial	154	3,347	Chanka	2021	Justice
focus_2007_quz	220	3,238	Collao	2007	Narrative
literatura	190	2,930	Chanka	1999	Culture
guia_collao	288	2,824	Collao	2015	Education
wikimedia	163	2,712	Collao	2021	Miscellaneous
docente	286	2,550	Chanka	2015	Education
convencion	115	2,548	Collao	1994	Agreement
yupaychaqa_ley	129	2,484	Chanka	2014	Norma
mikhunanchiskunamanta	127	1,925	Collao	2013	Social
tatoeba	428	1,778	Collao	2021	Miscellaneous
nanoquechua	92	1,431	Collao	2016	Culture
kallpa_qu	100	968	Collao	2019	Narrative
defensoria	60	882	Chanka	2021	Justice
yachay	62	756	Collao	2015	Culture
Total	**384,184**	**4,408,953**	-	-	-

Table 4: Details of each corpus included in the Southern Quechua corpus introduced.

Tokenization Approach		NER Class						
		B-LOC	B-ORG	B-PER	I-LOC	I-ORG	I-PER	O
BPE	True Positive	453	81	189	300	226	162	477
	False Positive	319	11	150	71	51	87	31
	False Negative	64	37	79	171	80	207	82
Norm. and BPE	True Positive	451	70	187	302	227	196	470
	False Positive	299	8	138	83	51	94	32
	False Negative	66	48	81	169	79	173	89
Norm. and PRPE	True Positive	449	79	187	306	227	186	471
	False Positive	304	14	135	95	53	74	28
	False Negative	68	39	81	165	79	183	88
Norm. and BPE-Guided	True Positive	453	71	176	299	222	156	466
	False Positive	294	16	164	93	57	113	28
	False Negative	64	47	92	172	84	213	93

Table 5: Breakdown of prediction results used to calculate weighted precision, recall, and F1 for the NER task .

POS Class		Algorithm			
		BPE	Norm. and BPE	Norm. and PRPE	Norm. and BPE-Guided
adj.	True Positive	253	259	262	235
	False Positive	98	106	92	96
	False Negative	143	137	134	160
verb	True Positive	764	760	761	744
	False Positive	77	86	72	98
	False Negative	78	82	81	72
pron.	True Positive	36	36	37	34
	False Positive	14	13	13	18
	False Negative	7	7	6	9
prep.	True Positive	0	0	0	0
	False Positive	0	1	0	0
	False Negative	1	1	1	1
adv.	True Positive	183	184	188	161
	False Positive	57	53	56	51
	False Negative	50	49	46	73
pron. indef.	True Positive	0	1	1	1
	False Positive	0	0	0	0
	False Negative	2	1	1	1
adv. interr.	True Positive	1	1	1	1
	False Positive	0	0	0	0
	False Negative	0	0	0	0
pron. interrog.	True Positive	8	7	8	7
	False Positive	5	2	5	2
	False Negative	2	3	2	3
num.	True Positive	0	0	0	0
	False Positive	0	0	2	3
	False Negative	5	5	5	5
conj.	True Positive	7	8	8	8
	False Positive	6	6	6	8
	False Negative	5	4	4	4
det.	True Positive	39	39	39	43
	False Positive	33	36	33	38
	False Negative	20	20	20	16
subj.	True Positive	1380	1376	1386	1380
	False Positive	138	124	131	193
	False Negative	112	115	107	113
interj.	True Positive	0	0	0	0
	False Positive	0	0	0	5
	False Negative	3	3	3	3

Table 6: Breakdown of prediction results used to calculate weighted precision, recall, and F1 for the POS task .