CS447 Literature Review: Natural Language Processing for Indigenous Languages of the Americas

Diana González Santillán,
dianag4@illinois.edu

November 26, 2021

Abstract

In this work we examine five articles that cover several aspects of the latest advances in Natural Language Processing (NLP) for a sample of indigenous languages of the Americas. We explore how each of these studies use existing resources (typically scarce) to develop tools and models for the aforementioned languages, we also review how well these tools and models perform, and describe the challenges faced by researchers when attempting to develop them.

1 Introduction

The specific question that we aim to answer in this literature review is: ”What Natural Language Processing (NLP) resources and models already exist for indigenous languages of the Americas, how well do they work, and what are the challenges faced when designing them?”

This paper is a short sample on the current state and challenges of NLP in low-resource, and many times underrepresented language groups, focusing on the indigenous languages of the Americas. Is is important to study this topic from a social equity standpoint, as well as from an academic interest point of view.

On the social equity side, the Americas are home to a wide range of linguistic families (comprising around 900 different languages) that are still spoken by millions of people (Mager et al., 2018). Many of these indigenous languages unfortunately face a risk of extinction, even though their existence increases the cultural and linguistic richness of the continent (Mager et al., 2018). It makes sense from a social and ethical standpoint to find a way to develop models that are inclusive of underrepresented languages to make sure that their speakers have access to the broad range of services that Natural Language Processing based tools has to offer, as well as to ensure that their cultures do not disappear.

On the academic interest side, the native languages of the Americas exhibit linguistic phenomena that are different from the most common languages usually studied in NLP. Therefore, studying them in this context will allow us to complement and develop more general NLP models that cover the broad diversity of languages and linguistic phenomena that exist in the world (Mager et al., 2018). In addition, many of these languages are low-resource languages, which also poses an interesting challenge for which the solution, if found, could be applied to other low-resource Artificial Intelligence problems.

2 Paper Summaries and Analysis

2.1 Challenges of Language Technologies for the Indigenous Languages of the Americas

2.1.1 Paper Summary

In this paper Mager et al. (2018) begin by stating the importance of studying indigenous languages of the Americas in the NLP field, and describe some of the outstanding linguistic characteristics that make the native American languages so linguistically diverse and unique. Mager et al. (2018) also highlight that due to the generalized deprivation of formal education in indigenous communities, many indigenous languages face a lack of orthographic normalization, which becomes a recurrent topic across the paper, and is considered to be one of the most prominent challenges in this field of study.

Mager et al. (2018) also mention that indigenous languages of the Americas tend to have low digital text production, as they are mostly oral and have many dialects across regions. This is a challenge since many NLP methods require vast amounts of corpora to achieve good performance. Despite this, Mager et al. (2018) still explore and enumerate the digital resources that do exist for some of the native languages spoken in the Americas (mainly Nahuatl, Wikarika, Shipibo Konibo, Guarani, and Quechua).

Besides listing the digital resources that exist for native languages of the Americas, Mager et al. (2018) also review several studies on morphology tasks and how these are applied to the Americas languages. Several examples of rule-based, semi-supervised, and unsupervised approaches on specific indigenous American languages are then enumerated and summarized. In this section, Mager et al. (2018) also describe some key characteristics of the America's languages that distinguish them from other languages commonly studied in NLP, as well as the morphology approaches different papers have taken to deal with this.

Furthermore, Mager et al. (2018) enlist studies in Machine Translation (MT), stressing their importance in the context of indigenous languages. In particular, Rule-Based MT (RBMT) approaches seem to be more suitable for low-resource languages since they do not require aligned parallel corpora. Nevertheless, RBMT models sometimes have shortcomings when attempting to translate complex constructions. An alternative approach is Statistical MT, which works better for complex constructions but highly depends on the amount of training data. In this regard, Mager et al. (2018) mention some techniques to reduce data sparseness, as well as some attempts at automatic data collection.

In addition to morphology tasks and MT, Mager et al. (2018) also mention some works on multilinguality and code-switching, both of which are common occurrences among communities that speak native American languages. Being able to detect code-switching would be useful when using the web as a source corpus, as well as in the field of quantitative linguistics. Furthermore, the paper includes a set of language tools that work on indigenous languages of the Americas to accomplish a variety of tasks such as speech synthesis and recognition, POS tagging, spell checking, language identification, and parsing.

Finally, Mager et al. (2018) discuss the increase in works related to NLP on indigenous languages of the Americas in recent years. Nevertheless, the developed technology on NLP for these languages is still not enough, as it only barely touches 35 out of the 900+ existing languages. Moreover, problems that are considered almost solved in languages like English still need to be adapted or started from scratch when it comes to native American languages.

2.1.2 Paper Analysis

This paper is certainly not exhaustive, and could be considered outdated. Nevertheless, it is still a good introduction to the literature on NLP in relation to the indigenous languages of the Americas. It enlists numerous tools, and resources that exist, and although it does not go into detail on the performance of each of them, it does go over a few of the main challenges faced by researchers that are still valid today.

In terms of the existing tools and models, the paper highlights that most of the works on NLP for indigenous languages of the Americas fall within two main categories, morphology tasks, and MT. Morphology is a logical category to study first, as it is a basic building block that must be developed before more higher level research can be performed. Meanwhile, studying MT also makes sense, as indigenous languages tend to coexist with other official languages in their countries. Other smaller categories such as multilinguality and code-switching are also noteworthy, expecially from a socio-linguistic point of view.

In terms of the challenges faced, the paper makes emphasis on the lack of orthographic normalization across the majority of studies reviewed. Another evident challenge for the study of NLP on native American languages is the scarce digital resources available as data, which is fundamental for good model performance. In addition, the few resources that do exist, are concentrated in just a handful of the hundreds of languages in the continent, which makes it challenging to study the whole spectrum of languages available.

2.2 Ayuuk-Spanish Neural Machine Translator

2.2.1 Paper Summary

In this paper Zacarías M. and Meza R. (2021) present the first neural machine translator system for the Ayuuk language (spoken in the state of Oaxaca in Mexico) into Spanish. The steps taken to create a low-resource parallel corpus between the two languages are described, as well as the specifications for the transformer neural architecture used to create the translator. Zacarías M. and Meza R. (2021) also stress the importance of creating machine translation models for native American languages since these technologies can provide indigenous people with access to knowledge and legal, medical, and financial services in their own language.

Zacarías M. and Meza R. (2021) first describe the nuances of the Ayuuk language which, besides having six different variants across Oaxacan communities, does not yet have a normalized orthography - with the number of consonants to be used being the main point of debate. There are two main positions in this argument, the “bodegueros” and the “petakeros” and Zacarías M. and Meza R. (2021) made sure to include both orthographies when normalizing texts to create their parallel corpus.

After normalizing orthography, Zacarías M. and Meza R. (2021) list the different resources in both languages that were used to create their parallel corpus. These resources included the Bible, the Mexican Constitution, a few short literary pieces, and phrases translated by one of the authors. Some of these resources were already aligned, and for the others Zacarías M. and Meza R. (2021) used the YASA automatic alignment tool. After aligning, they split the data into training, development, and testing sets in two different configurations: “strict” and “random”.

Next, Zacarías M. and Meza R. (2021) describe the Transformer Neural architecture used to create an encoder-decoder translator with two different configurations: one with 3 layers and embed size of 64, and one with 6 layers and embed size of 256. Both of these models were trained in a local server as well as in Colaboratory. Once the models were trained, experiments were made (5 experiments per data split version), and the results and learning curves of these experiments can also be found in the paper.

Furthermore, Zacarías M. and Meza R. (2021) dedicate a section to analyze the results of the experiments, confirming that translation between Ayuuk and Spanish is indeed possible, and “easier” from Spanish to Ayuuk than in the other direction. Zacarías M. and Meza R. (2021) also state that the model with more layers performed better overall, pointing out that this is not a trivial result, given that there was such a small amount of training data available.

Finally, Zacarías M. and Meza R. (2021) conclude that a standard model based on the Transformer architecture is way more promising than previous MT attempts on similar extremely low-resource problems. Zacarías M. and Meza R. (2021) also speculate that for future work they will focus on collecting more data, developing a deeper morphology analysis of the language, and improving orthographic normalization.

2.2.2 Paper Analysis

This paper focuses on Machine Translation, which seems to be one of the common topics when studying NLP for indigenous languages of the Americas. In particular, it describes the development of a specific first-of-its-kind Spanish-Ayuuk translator model. The performance of this model is evaluated as part of the study, and several challenges faced when developing it are addressed. The paper also mentions not only the academic but also the social reasons why developing translation models like this one is important.

On the performance front, the fact that such a translator is even possible is a relevant result given it was the first of its kind for this low-resource language. In comparing performance between the versions of the same model, it was determined that the Spanish to Ayuuk translation was ”easier” than the other way around, and that more layers in the model performed generally better. In comparing to other similar low-resource translation models, it was shown that using a transformer approach actually improved performance.

Regarding the challenges, it is not surprising to find that scarce data and lack of orthographic normalization were two of the main concerns. The paper also mentions that developing a translator for a new language with little to no NLP studies on morphology analysis poses a challenge. Therefore, it makes sense that Zacarías M. and Meza R. (2021) plan to find (or develop) such analyzers as well as better ways to gather more data before attempting to create an improved version of the translator model.

2.3 Peru is Multilingual, Its Machine Translation Should Be Too?

2.3.1 Paper Summary

In this paper Oncevay (2021) proposes the first multilingual translation model (many-to-Spanish and vice-versa) for indigenous languages spoken in Peru, specifically focusing on: Aymara, Ashaninka, Quechua, and Shipibo-Konibo. Oncevay (2021) stresses the importance of having multilingual Machine Translation (MT) models that better represent the rich diversity of languages in the country, many of which are considered endangered. In addition, Oncevay (2021) also states that using a multilingual approach can help mitigate the fact that all of these languages do not have enough resources for individual models to be developed.

After a brief description of each of the resources exploited to develop the translator model, Oncevay (2021) mentions two main challenges faced. First, that all four languages studied are highly agglutinative or polysynthetic - they usually express a large amount of information in just one word. Second, that initially the datasets were noisy and not cleaned, although this was later resolved. To continue, Oncevay (2021) explains how the data was split for training and testing, and describes how the model was evaluated using BLEU and chrF metrics. Next, a section on multilingual subword segmentation explains how words were processed to split affixes and preserve roots using a unigram language model, Oncevay (2021) also mentions that in this process all four languages were sampled with a uniform distribution. Moreover, Oncevay (2021) explains the procedure for the experiments conducted.

To continue, Oncevay (2021) decribes the methodology used to develop the translator model. Namely, a transformer-based model with default configuration in Marian NMT. Two models were pre-trained with a Spanish-English language-pair corpus in both directions, then, following established practices from other multilingual models, Oncevay (2021) fine-tuned the desired multilingual model. Furthermore, Oncevay (2021) explains how back-translated (BT) monolingual data was also used to enhance the model. However, it is mentioned that results using this data actually underperformed or did not converge, possibly because of the large amount of BT sentences versus human-translated sentences. This issue was alleviated by adding a special tag for the BT data.

In the next section, Oncevay (2021) analyzes the outcomes of the experiments. It is mentioned that another approach to alleviate the deteriorated performance of the models when using BT data could be considering more informed strategies for denoising, or performing online data selection. Oncevay (2021) also highlights that the multilingual model without BT data outperforms the other models in all languages except Quechua, which happens to be the “highest”-resource language out of the four. To continue with the analysis, Oncevay (2021) points out that in general there was no overfitting, with the exception of Spanish to Ayamara and Spanish to Quechua. In addition, Oncevay (2021) discusses that the metric used for scoring these models (BLEU) works at the word-level but other character-level metrics might need to be considered to better assess the highly agglutinative nature of the languages.

Oncevay (2021) concludes by stating that multilingual MT models can enhance the performance in truly low-resource languages, but it seems that all languages used need to have the same level of low-resources to be successful. Otherwise, we will obtain results like the ones for Quechua in this study, which have worse performance than the other languages.

2.3.2 Paper Analysis

This paper is another instance of Machine Translation research in regards to native American languages. In this case the study focuses on developing a multilingual translation model, and uses data that is not necessarily available in the target languages for pre-training, which is an interesting way to tackle the low-resource issue that makes it hard to work with many if not all of the indigenous languages of the Americas.

The model presented in this paper performed well in general, and better than pre-existing individual models for each of the same languages, except Quechua. Thus, when a language is extremely low-resource, it benefits from being modelled along with other similar low-resource languages, however, when a language has more resources available than the others in a multilingual model, it becomes negatively impacted.

Overall this paper poses novel approach to machine translation on indigenous languages of the Americas, evaluates interesting results, and tackles some of the challenges commonly encountered in this area of research. Nevertheless, the metrics used to evaluate its models might not cover all possible nuances of the highly agglutiative languages. Navive American languages in general are agglutinative so not having the correct metrics as a standard could be considered as another challenge on the field.

2.4 A Coreference Corpus and Mention-pair Baseline for Coreference Resolution in Quechua

2.4.1 Paper Summary

In this paper Pankratz (2021) makes two important contributions to the field of coreference resolution for low-resource languages by presenting the first coreference corpus to be developed for a Quechuan language (qxoRef), as well as a baseline mention-pair coreference resolution system developed for this corpus. In the introduction, Pankratz (2021) defines the coreference problem and outlines its importance in any NLP pipeline. To continue, Pankratz (2021) describes how there exist two main varieties of Quechua that differ lexically, morphologically, and orthographically: a smaller Quechua I and a much larger Quechua II. The corpus introduced in this paper, is based on Quechua I, which makes the task more challenging due to complexity and low resources.

Next, Pankratz (2021) gives a quick review of Quechua I grammar. Some characteristics they highlight is that the language is agglutinative and morphologically complex, and that it supports null arguments. According to Pankratz (2021), null arguments are especially tricky to encode in a coreference corpus, in particular when there is little training data and no syntactic annotation available. Therefore, null arguments were just ignored in qxoRef.

To gather data for this corpus, Pankratz (2021) used transcriptions of twelve recordings of stories belonging to a larger audio corpus of native Quechua speakers. Postprocessing was applied to these transcriptions to normalize orthography, unify the morphological analyses and glosses, and translate into English. In addition, problematic artefacts of speech data were removed, the stems were POS-tagged, sentences divided, and mentions manually annotated by Pankratz (2021).

Afterward, Pankratz (2021) moves on to describe the structure of the corpus. Mentions in qxoRef are split into two classes: nouns and pronouns, which are then split into a few more subcategories that are all described in detail. Next, Pankratz (2021) goes over two limitations. Namely, that the data had not been syntactically parsed to produce slots in the sentences where null arguments would be, and that annotating only nouns and pronouns does not involve as many degrees of freedom as other corpora with more granular classes have.

In the next section, Pankratz (2021) shifts to describe the mention-pair baseline coreference resolution system trained on the novel corpus. Given a pair of mentions - a candidate anaphor and a candidate antecedent - a binary random forest classifier is trained to predict whether that pair is coreferential. Pankratz (2021) poses that using binary random forests means that the task is simpler than deep learning, and less data is required. In addition, this approach can more easily tell us which features are important for establishing coreference in the available data which is helpful for conducting error analysis.

To continue, Pankratz (2021) goes into detail on how the model was trained, and enlists all the features that were considered for the classification task, as well as the ones that were not considered due to the lack of proper syntactic parsers and embeddings for the language. Pankratz (2021) also describes three heuristics that were used to create training datasets, and notes that in all cases, singleton mentions in the data were removed. Moreover, Pankratz (2021) also describes the method by which the test datasets were created. Finally, Pankratz (2021) explains how the trained classifiers were applied to the test data using the ”closest-first clustering” method.

In the last section of the paper, Pankratz (2021) outlines the evaluation and error analysis applied on the three different heuristics used to split training data. Furthermore, Pankratz (2021) compares and evaluates the results in detail, making conjectures on how the volume and characteristics of each dataset might have influenced in the final results. The general pattern for all three heuristics was high precision and low recall. In regards to shortcomings, Pankratz (2021) highlights that classifiers favoured string and morpheme similarity, and fell short when dealing with coreferential mentions that require grammatical or discourse based information. In addition, classifiers frequently failed to identify an antecedent for demonstrative pronouns. Pankratz (2021) points out that including null arguments in the data would possibly have yielded more accurate results.

Finally, Pankratz (2021) concludes by mentioning that for future work, the corpus should be improved by engaging multiple annotators (and computing inter-annotator agreement) to avoid biases, syntactically parsing sentences, and allowing null arguments. In addition, Pankratz (2021) stresses that feature representations and embeddings should be considered for future versions of the corpus. To close, Pankratz (2021) does not miss to remind us of the social and academic importance of improving basic NLP toolkits for low-resource languages.

2.4.2 Paper Analysis

This paper presents a tool for Quechua I that falls outside the general pattern of studies on morphology tasks or MT for indigenous languages. In this case, a coreference resulution model is presented, which falls somewhere above the lower level morphology tasks but below the upper level translation tasks in an NLP pipeline.

The study does evaluate the performance of the tool, however, it does not have much to compare with, as the tool is the first of its kind. Thus, the evaluations and comparisons are performed across different heuristics applied to the same data and model. It is interesting to see that the majority of shortcomings on this model are a consequence of the lack of lower-level resources and small amount of data available for the particular language, which seems to be a recurrent theme across all NLP studies for indigenous languages of the Americas.

Nevertheless, it is also interesting to note that in this case, the study did not run into any challenges due to orthographic normalization, given that Quechua is one of the few native American languages that actually has a standardized orthography.

2.5 A Finite-State Morphological Analyser for Paraguayan Guaraní

2.5.1 Paper Summary

In this paper Kuznetsova and Tyers (2021) present the first morphological analyser for Paraguayan Guaraní. The language is widely spoken in South America and one of the official languages of Paraguay, however, it still does not have a wide range of freely-available data and tools for NLP systems. Therefore, Kuznetsova and Tyers (2021) state that data-driven approaches for morphological analysis are hard to apply, and instead develop a technique relying on formal linguistic descriptions and finite-state transducers.

In the first section, Kuznetsova and Tyers (2021) outline many of the particularities of Paraguayan Guaraní that are relevant when considering the task for morphological analysis. In addition, Kuznetsova and Tyers (2021) highlight that Paraguayan Guaraní only just recently standardized their orthography so writing standards in any data published before 2018 vary significantly. Furthermore, Kuznetsova and Tyers (2021) mention that most of the existing computational resources for Guaraní are not always reliable and lack many words. In terms of prior work, Kuznetsova and Tyers (2021) make emphasis on an existing finite-state morphological analyzer, *ParaMorfo*, which is very close to the analyzer developed by Kuznetsova and Tyers (2021) but focuses more on the form generation rather than morphological analysis.

Kuznetsova and Tyers (2021) then move on to describe the details of their open-source FST-based analyzer development. The two level transducer uses two formalisms that follow the HFST platform conventions, *lexc* - morpheme combinatorics or morhpotactis, and *twol* - phonological rules or morphophonology. On the morphotactics side, Kuznetsova and Tyers (2021) describe how stems in Guaraní are ambiguous, as they could refer to nominal (nouns, adjectives, adverbs) or verbal classes which causes the same stems to appear in various basic lexicons. Then, on the morphophonology side, Kuznetsova and Tyers (2021) briefly go over 30 phonological rules of Guaraní that are included in their model.

To continue, Kuznetsova and Tyers (2021) outline the evaluation process for their morphological analyzer. They start by estimating its naive coverage metric (the ratio of tokens that receive at least one morphological analysis to the total number of tokens in the corpus) and comparing it to that of the *ParaMorfo* system, ensuring fairness by dropping from the testing data tokens that *ParaMorfo* does not recognize. It is concluded that Kuznetsova and Tyers (2021)'s analyzer still performed significantly better than *ParaMorfo*, probably because it is able to cover Spanish barbarisms as well as recognize proper names, and it is more flexible with orthographic variation.

Besides comparing to *ParaMorfo*, Kuznetsova and Tyers (2021) point out that conventional quality metrics such as precision, recall, and F-measure, and the average ambiguity rate (average number of analyses given by the transducer per word) were also evaluated. The results of the first three metrics point to the likelihood of a word being analysed correctly (if it is analysed) as being fairly high. Furthermore, the result for the average ambiguity rate means that Guaraní is moderately polysynthetic as compared to other languages. Finally, Kuznetsova and Tyers (2021) grant that their analyser did not correctly solve the case of pronouns serving as both possessive and personal markers, as the data presumes only personal pronouns exist.

2.5.2 Paper Analysis

This paper introduces a morhpological analyzer, which falls into one of the two main categories of studies on indigenous languages we have seen as a pattern so far. The paper is concise and goes to the point of presenting and evaluating the analyzer developed. Given this, it does not go into much detail analyzing the challenges faced nor does it discuss possibilities for future work.

The evaluation of the tool shows that it has favorable results compared to previous similar tools, and highlights that these results are due to being able to deal with barbarisms and orthographic variation, which is a recurrent challenge across studies of this type. The lack of resources for this language is briefly addressed as well, and is evidently also a challenge, which is worked around by using finite state transducers instead of more data intensive approaches.

This morphological analyzer is important because it is the first of its kind for the Guaraní language, and it provides a fundamental building block for developing more complex, higher-level, NLP tools for the language

3 Discussion and Conclusion

The question asked at the beginning of this review was ”What NLP resources and models already exist for indigenous languages of the Americas, how well do they work, and what are the challenges faced when designing them?”. Although this is not an exhaustive literature review, we can still see there is a pattern across the reviewed works in how to answer the three parts of this question.

When it comes to what resources and models already exist, we see that several tools and models have been created, especially in the last couple of years. Most of these new tools are first of their kind for the particular native American language they apply for, and they mostly fall into the categories of morphological analyzers and machine translators. Nevertheless, there are also other types of tools being developed, such as co-reference analyzers, and studies on multilinguality and code-switching. As to the resources available, most, if not all, indigenous languages of the Americas seem to be in extremely low-resource conditions, and many of the tools developed so far also had to develop their own corpora or workarounds to be able to train and test their models.

In regards to how well all of these existing resources and tools work, since they are mostly first of their kind, it is difficult to compare their results to results of other tools with the exact same objective. Despite this, all of the tools and models reviewed in this paper seem to outperform similar tools in other languages, and show promising results based on the thresholds set for each study considering the challenges faced. However, perhaps these papers are just setting the low bar for future research, as it is easy to perform well when presenting a first-of-its-kind tool. It is also noteworthy that some of the metrics used to evaluate the studies reviewed in this paper might have been designed to evaluate NLP tools in English, and might not be the best at evaluating languages have a very different structure and grammar.

In terms of the challenges faced, we can also find a notorious pattern showing that the main two issues with NLP studies on indigenous languages of the Americas are first, the low-resource setting most of these languages are in, and second, the lack of orthographic and grammatical normalization for the written versions of these languages. In several instances it is also mentioned that another challenge is the lack of lower-level NLP tools, such as syntactic and grammatical parsers, that would make higher-level tasks such as translation and co-reference parsing much more straightforward.

In conclusion, although a good amount of research and development has already been done in the topic of Natural Language Processing for Indigenous Languages of the Americas, there is still a long way to go before research on these languages can reach the level of depth that has been reached for other more popular languages. Some NLP problems that have already been solved for languages like English need to be re-evaluated and many times started from scratch to be able to solve them for languages as complex as those of the Americas. Regardless, it is still important to attempt to solve these problems. From an academic point of view, to make research more thorough and explore interesting computational and linguistic challenges. As well as from a social point of view, to honor, and preserve the cultural heritage indigenous languages of the Americas bring to the world, and be inclusive of the speakers of these languages so they are able to benefit from the technological advancements and services that Natural Language processing has to offer.

References

Anastasia Kuznetsova and Francis Tyers. 2021. A finite-state morphological analyser for paraguayan guaraní. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, Online. ACL Anthology.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza. 2018. Challenges of language technologies for the indigenous languages of the americas. In *arXiv:1806.04291*, Online. arXiv.

Arturo Oncevay. 2021. Peru is multilingual, its machine translation should be too? In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, Online. ACL Anthology.

Elizabeth Pankratz. 2021. qxoref 1.0: A coreference corpus and mention-pair baseline for coreference resolution in conchucos quechua. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, Online. ACL Anthology.

Delfino Zacarías M. and Ivan Vladimir Meza R. 2021. Ayuuk-spanish neural machine translator. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, Online. ACL Anthology.