Language Identification of Philippine Creole Spanish: Discriminating Chavacano From Related Languages

Aileen Joan Vicente
De La Salle University
Philippines
aileen_vicente@dlsu.edu.ph

Charibeth Cheng
De La Salle University
Philippines
charibeth.cheng@dlsu.edu.ph

Abstract

Chavacano is a Spanish Creole widely spoken in the southern regions of the Philippines. It is one of the many Philippine languages yet to be studied computationally. This paper presents the development of a language identification model of Chavacano to distinguish it from languages that influenced its creolization using character-level Convolutional Neural Networks (CNN). Unlike studies that discriminated similar languages based on geographical proximity, this paper reports a similarity based on a language's creolization. We established the similarity of Chavacano and its related languages, Spanish, Portuguese, Cebuano, and Hiligaynon, from historical accounts and lexical similarity based on the number of common words in the corpus for all languages. We report an accuracy of 93% for the model generated from a CNN using ten filters with a filter width of 5. The training experiments reveal that increasing the filter width, number of filters, or training epochs is unnecessary even if the accuracy increases because the generated models present irregular learning behavior or may have already been overfitted. This study also demonstrates that the character features extracted from CNN, similar to n-grams, are sufficient in identifying Chavacano. Future work on the language identification of Chavacano includes improving classification accuracy, especially for short or code-switched texts for practical applications such as social media sensors for disaster response and management.

1 Introduction

Language Identification (LI) is the task of deciding which natural language a particular text is written in. The research in this field aims to mimic the ability of humans to recognize these languages. LI enables many natural language applications and language processing (NLP) tasks. For example, automatic machine translation applications must identify the text's language before translating it into English. It can be used for document collections where the languages of the documents are unknown beforehand (Jauhiainen et al., 2019), such as in the case of crawling the web as part of corpus-building.

Many LI systems and studies target English and other major languages spoken worldwide. It is especially understandable since large repositories of language texts exist for these languages. There are also initiatives to identify low-resource languages such as Uralic languages (Jauhiainen et al., 2020) and Austronesian languages (Dunn and Nijhof, 2022). However, many other low-resource languages do not have enough digital resources for extensive research. While LI is generally considered a solved task, the work on LI for low-resource languages persists due to the widespread use of the Internet and the development of applications based on natural language understanding, such as chatbots. Selamat and Akosu (2016) argued that the inability to identify a language makes the language invisible in any multilingual environment, such as in the case of Chavacano, the Philippines' Creole Spanish.

Chavacano is one of those under-researched, low-resource languages. Websites with automatic translations identify Chavacano as Spanish, given the former's similarity with the latter. Chavacano's lexicon is predominantly Spanish (Lipski and Santoro, 2007) but with orthographic shifts.

Languages can differ in many ways. They may use different sounds, other writing systems, different vocabulary, or put words together to form a sentence differently. For similar languages, however, such as language variants and dialects, discriminating between them remains challenging (Zampieri et al., 2014) and is one of the bottlenecks of state-of-the-art language identification systems.

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 188–196
June 20, 2024 ©2024 Association for Computational Linguistics

This paper reports the language identification of Philippine Creole Spanish. Unlike the similar languages investigated in the Discriminating between Similar Languages (DSL) shared tasks, whose language similarities are mostly due to geographic proximity, this study investigates the identification of a Creole, i.e., Chavacano, among its related languages.

This study brings forward the unique characteristics of Philippine Creole Spanish (PCS) as an amalgamation of foreign and native languages. In the case of Chavacano, it is the complex intermixing of Spanish, Portuguese, Cebuano, and Hiligaynon during centuries of colonization, migration, and trade. The linguistic features of Chavacano that combine elements of multiple language sources make it a linguistically rich and unique variety.

The language identification of Chavacano is expressed as a character-level sentence classification that discriminates among similar, related languages and where the languages are considered as the target classes.

The remainder of this paper is organized as follows. Section 2 presents the linguistic properties of Chavacano and its similarities with related languages. Section 3 introduces related works implementing CNN for language identification. Section 4 gives a detailed overview of the steps to build the language identification model. In particular, Section 4.2 provides an overview of char-CNN, the character-level Convolutional Neural Network used to train the model. In Section 5, we report and analyze our experimental results, while Section 6 concludes this paper and gives some directions for future research.

2 Chavacano: Philippine Creole Spanish

The Philippine Creole Spanish, collectively known as Chavacano, comprises three major dialects spoken in Ternate, Cavite, and Zamboanga (Lipski, 2001). Both the Ternate and Cavite dialects are classified as the Manila Bay PCS. Ternateño was the oldest Spanish-based Creole in the Philippines, and Caviteño was an off-shoot. Zamboangueño, on the other hand, comprises the largest group of Chavacano speakers in Zamboan a City and neighboring towns and cities in Mindanao. In this study, we refer to the variant Zamboangueño, as it is the only thriving variant. Aside from the population of speakers, Zamboangueño is actively used in blogs, news, and social media that can be used as digital resources.

The formation of Chavacano in Zamboanga resulted from historical and cultural interactions in the Philippines during the Spanish colonial period from 1565 to 1898. Chavacano belongs to the Creole family of languages of Spanish descent (Eberhard et al., 2023).

The language started to develop during the Spanish garrison in Zamboanga, beginning with the absorption of grammatical and lexical structures from Manila Bay PCS in the 18th century. Manila Bay PCS is said to have been influenced by the Portuguese language (University of Hawai'i Press, 1975; Lipski, 2001). Ilonggo or Hiligaynon later influenced Chavacano as Iloilo became a stopover for ships from Manila to Zamboanga. Later in the 20th century, immigration from the Central Visayan region to southwest Mindanao added some Visayan or Cebuano items to the language. Given this history, Chavacano is described as a "contact vernacular that has undergone numerous remakings by an ever-changing population that has never given up their native languages" (Lipski, 1992). It is easy to see that Chavacano's words are predominantly Spanish, but an inspection of usage tells us that they are not entirely Spanish.

Over three centuries of Philippine history influenced the morphology, grammar, and syntax of Chavacano (Lipski and Santoro, 2007). It has retained its Austronesian foundation, evidenced by the Verb-Subject-Object word order, with many alternative possibilities (Lipski, 1992). The Philippine languages belong to the Austronesian language family. This contrasts Spanish's Subject-Verb-Object word order (Lee, 2017).

The lexicon of Chavacano is largely Spanish (Lipski and Santoro, 2007) but with orthographic shifts. It has experienced several stages of relexification to include lexical items of Philippine origin from regional Visayan (Cebuano), Ilonggo (Hiligaynon), and occasionally Tagalog (Lipski, 2001). It has also adopted a heavy English lexical transfer (Lipski, 1992) over time.

Chavacano words are spelled using the alphabet of the word's traced etymology (DepEd-IX, 2016). For example, the Spanish-derived words *zacate* (grass) and *mañana* (tomorrow) are spelled using the Spanish alphabet, the *Abecedario*. In contrast, the Chavacano words of local origin, like *kanila* (them) and *kanamon* (us), are spelled using the Philippine alphabet system. The letter *r* in the Spanish verbs like *comer* (to eat), *bailar* (to dance) are dropped in Chavacano, i.e., *come*, *baila*. In general, Chavacano words are spelled the way they are pronounced. It is also interesting to note that the Spanish writing utilizes diacritics that are not necessarily applied in Chavacano.

In summary, Chavacano began as a hybrid pan-Philippine contact language whose Spanish items had already been filtered through Philippine languages and which was, therefore, a Philippine language in the structural sense at every point of its existence (Lipski, 2001).

3 Related Works

Jauhiainen et al. (2019) assert that from a computational perspective, the algorithms and features used to discriminate between languages, language varieties, and dialects are identical. Hence, the choice of features and algorithms depends on the researcher and the data used for the study.

Both discriminative and generative algorithms have been explored in more recent LI studies. Hidden Markov Models and Latent Dirichlet Allocation are the common generative methods used. Decision trees, support vector machines, neural networks, and ensembles are widely used discriminative models.

Characters are the building blocks of a language's writing system. Although most languages follow an alphabetic system, the languages still differ in character combinations and orthography. Hence, characters and their combinations have been widely used in LI.

An example of character combinations is n-grams. Character n-grams are widely used character sequences that may capture a language's orthography (Simões et al., 2014). Character n-grams are sequences (consecutive or overlapping) of characters of length *n*. The frequency of these n-grams has been used as feature vectors for most LI research involving discriminative methods.

Using CNN for LI is seen as a means of automatically extracting character features from text for classification. Zhang et al. (2015) was among the first to introduce character-level CNN for text classification. In this case, text is seen as a kind of raw signal at the character level where CNN extracts features (Zhang et al., 2015; Kim et al., 2016). The successful application of Zhang et al. (2015) and Kim et al. (2016) also sparked interest in CNN for LI. Guggilla (2016), Belinkov and Glass (2016), Jaech et al. (2016b), Jaech et al. (2016a), Ali (2018a), Ali (2018b), Chung et al. (2019) are among those who have successfully implemented CNN for LI. It has grown in acceptance in LI because it eliminates the need to extract or handcraft features separately, such as feature engineering.

4 Methodology

4.1 Data Preparation

The corpus used in the study is mixed-domain. The monolingual Hiligaynon and Cebuano sentences were taken from the PH-MNMT corpus (Coronia, 2022), which consists of web-scraped articles and bible translations. The Spanish and Portuguese sentences were mainly taken from the DSL Corpus Collection (Tan et al., 2014), which consists of news articles. Additional sentences for Spanish and Portuguese were taken from Bible translations as well.

On the other hand, the Chavacano sentences were collected from print sources (de Saint Exupéry (Author) and De Los Reyes (Translator), 2018) and online sources (Herrera; Zamboanga News Online; Wycliffe Bible Translators, Inc.).

The corpus contains 107,500 sentences with 21,500 sentences for each language (Table 1).

The raw sentences used in the corpus are made available at `https://github.com/ajvicente/cbk-li`.

The Spanish and Portuguese sentences were tokenized using tokenizers specific to the language. Cebuano and Hiligaynon, on the other hand, were tokenized using English-based tokenizers. Punctuation and numerical literals were later removed

Language	Source Domains	No. of Sentences	Traning Data	Testing Data	Validation Data
Chavacano	bible translations, blogs, book, feature articles	21,500	18,000	2,000	1,500
Cebuano	bible translations, web-scraped documents	21,500	18,000	2,000	1,500
Hiligaynon	bible translations, web-scraped documents	21,500	18,000	2,000	1,500
Spanish	bible translations, news articles	21,500	18,000	2,000	1,500
Portuguese	bible translations, news articles	21,500	18,000	2,000	1,500
		107,500	90,000	10,000	7,500

Table 1: Chavacano and Related Languages Corpus

from the data set. The texts were converted to lowercase after all unnecessary characters had been removed. The alphabet of the corpus contained 46 characters. Digraphs such as *ch*, *ng*, *rr*, and *lh* are counted as single characters. Characters with diacritics are also counted separately.

Language	Unique Words	Overlap Words with Chavacano	Unique Characters	Sentence Length (Max Characters)
Chavacano	5,740		36	424
Cebuano	26,486	981	27	639
Hiligaynon	20,832	941	27	574
Spanish	50,836	2,580	42	3,654
Portuguese	38,790	1,357	45	4,846

Table 2: Corpus Statistics

There are 5,740 unique Chavacano words in the corpus. Of these, 44.95% overlap with Spanish, 23.64% with Portuguese, 17.09% with Cebuano, and 16.39% with Hiligaynon (Table 2). Most of the shared words or overlaps are content words.

The small number of unique words in the Chavacano corpus is due to shorter sentence fragments in Chavacano and because most of the sentence fragments in the dataset were sourced from bible translations. Unlike the Cebuano, Hiligaynon, Spanish, and Portuguese datasets were primarily sourced from news articles and web texts covering more topics than bible translations. Hence, there is a greater variety of words in the related languages.

4.1.1 Character Encoding

The characters for each word in the corpus are sequentially encoded as in the work of Zhang et al. (2015). Encoding is based on an alphabet dictionary of size $m = 47$ that consists of the 46 common alphabet characters in the corpus and the space as the word delimiter. Each character is then quantized using 1-of-m encoding (or one-hot encoding). A fixed sentence length of $l = 1000$ characters is set. This value was empirically identified to cover all the words in the Chavacano sentence fragments. Shorter sentences are padded, while longer sentences (especially for Spanish and Portuguese) are truncated.

The labels are similarly one-hot encoded over five language classes: Chavacano, Cebuano, Hiligaynon, Spanish, and Portuguese.

4.1.2 Data Split

Training, Validation, and Test sets were extracted from the corpus using stratified sampling to ensure that all language classes are represented proportionally in each data set. 18,000 sentences per language are used for training, 2,000 for validation, and 1,500 for testing.

4.2 charCNN: Character-based Convolutional Neural Network

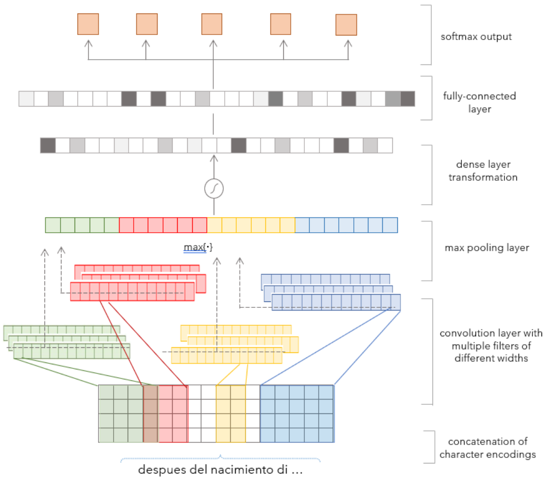
softmax output
fully-connected layer
dense layer transformation
max(·)
max pooling layer
convolution layer with multiple filters of different widths
concatenation of character encodings
despues del nacimiento di ...

Figure 1: charCNN Network Architecture adapted from Kim et al. (2016)

Following the work of Kim et al. (2016) and Zhang et al. (2015), a simple convolutional neural network was used to extract features from the training data and then fed to a dense layer for classification. Figure 1 illustrates the neural network architecture.

4.2.1 Convolution Layer

Based on Kim et al. (2016), a 2D convolution is applied between the input sentence $\mathbf{C}^s$ and a filter $\mathbf{H}$ $\in \mathbf{R}^{m \times w}$ where the filter width $w \in \{2, 3, 4, 5, 6\}$. With each filter, a feature vector $f^s \in \mathbf{R}^{(l-w)+1}$ is generated where the $i - th$ element of f is given by:

$$f^s(i) = \langle \mathbf{C}^s[*, i : i + w - 1]\mathbf{H} \rangle \qquad (1)$$

where $\langle \mathbf{A}, \mathbf{B} \rangle = Tr(\mathbf{AB}^T)$ is the Frobenius inner product.

Characters, as used in the study, correspond to signals in images, videos, and sounds (Zhang et al., 2015) that are typical inputs in CNN-based tasks.

4.2.2 Pooling Layer

The maximum value in f^s is extracted at the pooling layer as the feature corresponding to the filter $\mathbf{H}$ when applied to the sentence $\mathbf{C}^s$. According to Kim et al. (2016), in this process, the filter essentially picks out a character n-gram whose size of the n-gram corresponds to the filter width.

Given that multiple filters h are used in the study, then the representation of the input sentence is a concatenation of max pooling layers in the form $y^s = [y_1^s, ..., y_h^s]$.

A bias is added, and a non-linear transformation (tanh) is applied.

4.2.3 Dense Layer

A dense layer of 512 units followed by a dropout at 0.5 is added to the convolutional network before concluding with a softmax layer of 5 units to represent each of the five language classes. The categorical cross-entropy loss is used to fit the model. The model is optimized with Adam optimizer using a learning rate of 0.001.

4.3 Model Evaluation

Loss and accuracy metrics are collected during training (validation) and testing to evaluate the model's performance. The validation step during training uses the validation dataset to assess the model's performance during training. Model testing is performed after training using unseen data to simulate real-world scenarios. Ideally, the accuracy and loss values during validation and testing should be close enough to ascertain that the model does not overfit or underfit the data. The use of overall accuracy in this study is sufficient, given that the data is balanced for all language classes.

Several experiments that involved changes in the number of filters and combining filter widths are also conducted to arrive at optimized network parameters.

5 Results

Various training configurations using the number of filters (5, 10, and 15), range and combination of filter widths (2, 3, 4, 5, and 6), and number of epochs (10, 20, and 30) were experimented on in this study. The following sections report the result of such experiments and insights from the language identification modeling of Chavacano.

5.1 Experiments

The results of the experiments on various training configurations based on the number of filters, filter widths, and epochs show that the accuracy of the model naturally increases with increasing number of filters, filter widths, and epochs, as shown in Figure 2 and Figure 3.

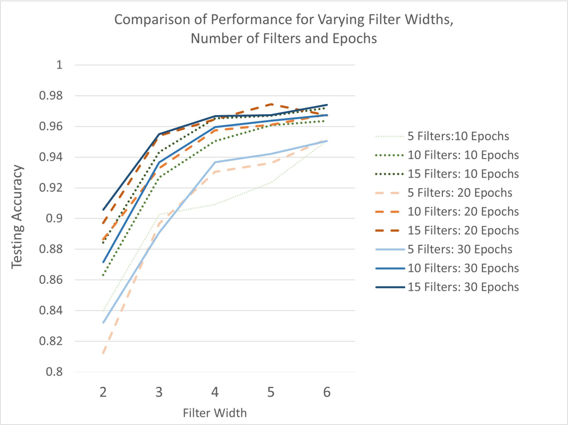

Figure 2: Comparison of Model Performances for Varying Filter Widths

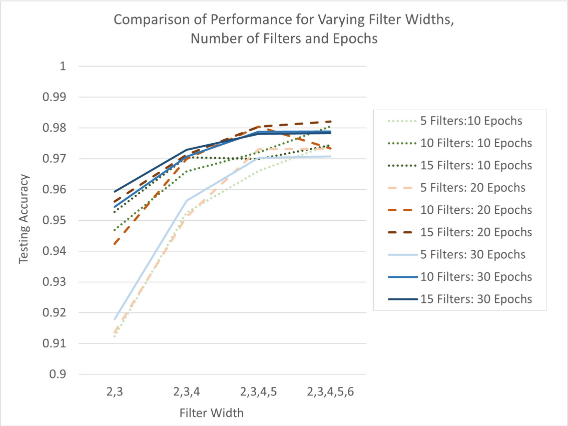

Figure 3: Comparison of Model Performances for Combined Filter Widths

The comparison in Figure 2 also shows that there is generally a sharp increase in performance using filter width 2 to 4 with the increasing number of filters and epochs, after which a slight and steady increase in the performance is observed except for the degradation of performance of the model learned at 15 filters and 20 epochs.

On the other hand, the combined filter widths in Figure 3 show similar behavior in the increase in accuracy until the combined filter widths of 2, 3, 4, and 5.

Figures 2 and 3 show that increasing the number of filters and the number of times these are seen during training does not necessarily contribute to a better model.

In the same way, a comparison of training and validation losses also reveals that although increasing the number of filters and the number of epochs increases validation accuracy, the model's training performance seemed irregular, as shown in an example in Figure 4.

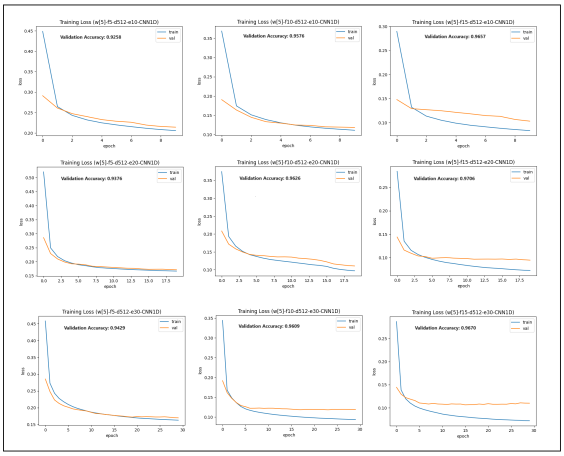

Figure 4: Comparison of Training/Validation Losses for Filter Width = 5.

The comparison shows that the divergence in the training and validation losses increases as the number of filters and epochs increases. This behavior indicates that the models may have already picked up noise in the data and overfit.

Finally, based on the model accuracy and variance of training and validation losses, the model generated using 10 filters with a filter width of 5 and trained in 20 epochs, earning a validation accuracy of 0.9376, is chosen as the best model among all training configurations.

5.2 Error Analysis

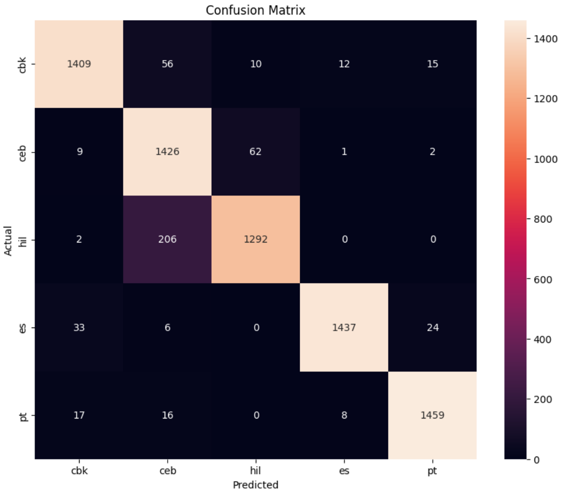

Figure 5: Confusion Matrix based on Model Testing

The confusion matrix in Figure 5 reveals that Chavacano can be confused with Cebuano, Hiligaynon, Spanish, and Portuguese. The related languages are also often mistaken for Chavacano.

It is observed that Hiligaynon and Cebuano, both local languages, are mostly confused with each other and that Hiligaynon is only confused with Cebuano and rarely Chavacano.

On the other hand, Chavacano is mostly confused with Cebuano, followed by Portuguese, Spanish, and Hiligaynon. Interestingly, Chavacano exhibits a greater overlap with Spanish and Portuguese when compared to Cebuano and Hiligaynon. Yet, Chavacano is mostly confused with the local language, Cebuano. This behavior may be attributed to Chavacano's orthography. Despite following the Spanish's Abecedario (DepEd-IX, 2016), Chavacano does not use many of the diacritics used by Spanish and Portuguese in writing.

The model confuses Spanish and Portuguese with Chavacano more than the local languages. In the case of Spanish, 20 of 33 (61%) misclassifications do not contain diacritics, and the rest of the 13 sentences only contained at most three characters with diacritics. For Portuguese, all 17 sentences that are misclassified did not contain diacritics.

The error analysis also revealed that 63% (35 of 56) of the Chavacano sentences misclassified as Cebuano were single-word sentence fragments. The longest misclassified sentence consists of 11 words. This result indicates that the model may be unable to correctly classify short sentences, significantly since most words overlap with other languages. Language identification involving short texts continues to be a challenging task for many languages (Jaech et al., 2016b; Jauhiainen et al., 2019).

The misclassification of Chavacano to Hiligaynon, Spanish, and Portuguese also share the same observation, albeit not as short as the Cebuano misclassifications. All misclassified sentences fall within less than 30% of the maximum number of words in the language's corpus.

6 Conclusions and Recommendations for Future Work

6.1 Conclusion

The experiments show that the language identification of Chavacano does not require a complex and deep CNN network. The model can already learn to discriminate the language from among its related languages using 10 filters with a filter width of 5. The hyperparameter search reveals that because the related languages share common characters to a large extent, it is vulnerable to overfitting. With the performance at 93%, the model can be used in the future to develop web applications to collect Chavacano documents.

This study demonstrates the viability of character features, specifically those generated by a convolutional neural network, to identify related languages. Instead of manually extracting n-gram features, this study demonstrates an end-to-end system of training a language identification model using neural networks.

The study also gleaned the orthographical similarities between Chavacano and Cebuano despite the latter being predominantly Spanish in cognates, although further studies need to be undertaken to establish this relatedness. Diacritics was also considered a contributing factor in discriminating Chavacano from Spanish and Portuguese.

6.2 Recommendations for Future Work

This paper presents a benchmark study for Chavacano LI that can be used as a baseline for future works. Further experimentation is recommended, including using other learning algorithms, such as SVM, or deep learning models, such as Transformers. In addition, the study uses mixed domains in training. The effect of the dataset domain in training needs to be experimented as this has been one of the issues in discriminating similar languages.

This preliminary work on Chavacano opens many other opportunities to understand and document Chavacano computationally and study Creole languages. The next step of this project is to implement the network design to discriminate Chavacano in natural settings, i.e., no preprocessing and within the context of multilingual documents. Based on the results, the language identification study can be extended to improve the classification of Chavacano in shorter, maybe code-switched, sentences such as those coming from Tweets to be used for practical applications such as social media sensors for disaster monitoring and management or more natural translation from code-switched sentences.

Limitations

While most language identification of related languages worked on dialects or variants, this study is limited to the related languages of Creole. The similarity is based on the languages' lexical, syntactical, and morphological influence on Chavacano. Another limitation is using CNN as the only model experimented with in the study. Experiments with other models to improve LI for Chavacano are encouraged as future works.

References

Mohamed Ali. 2018a. Character level convolutional neural network for Arabic dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 122–127, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mohamed Ali. 2018b. Character level convolutional neural network for German dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 172–177, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2016. A character-level convolutional neural network for distinguishing similar languages and dialects. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 145–152, Osaka, Japan. The COLING 2016 Organizing Committee.

Tonglee Chung, Bin Xu, Yongbin Liu, Chunping Ouyang, Siliang Li, and Lingyun Luo. 2019. Empirical study on character level neural network classifier for chinese text. *Engineering Applications of Artificial Intelligence*, 80:1–7.

Jeremy Dale Coronia. 2022. ph-mnmt-dataset. `https://huggingface.co/datasets/ecridale/ph-mnmt-dataset`. Accessed February 2023.

Antoine de Saint Exupéry (Author) and Robin De Los Reyes (Translator). 2018. *El Principe Niño: Der kleine Prinz - Zamboangueño Chabacano*, Creole edition. Edition Tintenfaß, Neckarsteinach, Germany.

DepEd-IX. 2016. *Zamboanga Chavacano Orthography*. Local Government of Zamboanga City: Philippines.

Jonathan Dunn and Wikke Nijhof. 2022. Language identification for Austronesian Languages. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 6530–6539, Marseille, France. European Language Resources Association (ELRA).

David Eberhard, Gary Simons, and Charles Fenning, editors. 2023. *Ethnologue: Languages of the World*, 26th edition. SIL International, Dallas, Texas.

Chinnappa Guggilla. 2016. Discrimination between similar languages, varieties and dialects using CNN- and LSTM-based deep neural networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 185–194, Osaka, Japan. The COLING 2016 Organizing Committee.

Jerome Herrera. Bien chabacano. `https://bienchabacano.blogspot.com/`. Accessed February 2023.

Aaron Jaech, George Mulcaire, Shobhit Hathi, Mari Ostendorf, and Noah A. Smith. 2016a. Hierarchical Character-Word Models for Language Identification. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 84–93, Austin, TX, USA. Association for Computational Linguistics.

Aaron Jaech, George Mulcaire, Mari Ostendorf, and Noah A. Smith. 2016b. A Neural Model for Language Identification in Code-Switched Tweets. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 60–64, Austin, Texas. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, Niko Partanen, and Krister Lindén. 2020. Uralic language identification (ULI) 2020 shared task dataset and the wanca 2017 corpora. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 173–185, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65(1):675–682.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2741–2749. AAAI Press.

James F. Lee. 2017. Word order and linguistic factors in the second language processing of spanish passive sentences. *Hispania*, 100(4):580–595.

John Lipski. 2001. The place of Chabacano in the Philippine linguistic profile. *Sociolinguistic Studies*, 2(2):119–163.

John Lipski and Maurizio Santoro. 2007. Zamboangueño creole spanish. In John Holm and Peter Patrick, editors, *Comparative creole syntax. Parallel outlines of 18 creole grammars*, Westminster Creolistics Series 7, pages 373–398. Battlebridge, London. Much information is based on Forman (1972).

John M. Lipski. 1992. New thoughts on the origins of zamboangueño (philippine creole spanish). *Language Sciences*, 14(3):197–231.

Ali Selamat and Nicholas Akosu. 2016. Word-length algorithm for language identification of under-resourced languages. *Journal of King Saud University - Computer and Information Sciences*, 28(4):457–469.

Alberto Simões, José João Almeida, and Simon D. Byers. 2014. Language Identification: a Neural Network Approach. In *3rd Symposium on Languages, Applications and Technologies*, volume 38 of *OpenAccess Series in Informatics (OASIcs)*, pages 251–265, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15, Reykjavik, Iceland.

University of Hawai'i Press. 1975. Chabacano (Philippine Creole Spanish). *Oceanic Linguistics Special Publications*, (14):210–216.

Wycliffe Bible Translators, Inc. El nuevo testamento. https://worldbibles.org/language_detail/eng/cbk/Chavacano. Accessed February 2023.

Zamboanga News Online. Comentarios desde zamboanga. http://comentariosdesdezamboanga.blogspot.com/. Accessed February 2023.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, pages 649–657. Curran Associates, Inc.