Bridging Ibaloi with NLP: Developing an LLM-based Translator using a Corpora-based Ibaloi Lexicon

Alcazar, Xymond Louisse
Saint Louis University
Baguio City
2227226@slu.edu.ph

Collo, John Henrich
Saint Louis University
Baguio City
2235110@slu.edu.ph

Lachica, Rafael
Saint Louis University
Baguio City
2195465@slu.edu.ph

Lucban, Prince John Louie
Saint Louis University
Baguio City
2225254@slu.edu.ph

Navarro, Josiah Ezra
Saint Louis University
Baguio City
2233059@slu.edu.ph

Olanos, Cheni Lei
Saint Louis University
Baguio City
2212637@slu.edu.ph

Yuen, Ka Hang Christian
Saint Louis University
Baguio City
2214959@slu.edu.ph

ABSTRACT

This study proposed an attempt to further enrich understanding and usage of the Ibaloi language using Natural Language Processing, precisely, the machine translation and language inference part. A lexicon is created using both secondary sources over the internet and a first-hand list of sentences that was interpreted and written by Ibaloi users. Doing so, the resources aim to provide insights on integration of NLP to languages and dialects so as to preserve them through usage.

Keywords

Ibaloi; Lexicon; NLP

1. INTRODUCTION

The Cordillera Administrative Region (CAR) of the Philippines is home to the largest concentration of Indigenous Peoples (IPs) in the country. Among its major ethnolinguistic groups are the Tinguian of Abra, Isneg of Apayao, Ibaloi of Baguio City and southern Benguet, Kankanaey of northern Benguet and southern Mountain Province, Ifugao of the Ifugao province, Bontok of northern Mountain Province, and Kalinga of the Kalinga province [1]. The Ibaloi language, being one of the most spoken languages in the Benguet province, plays a significant role in preserving the cultural identity and traditions of the Ibaloi people.

However, despite the presence of online Ibaloi dictionaries and references, the language remains underrepresented in technological applications, such as the need for a resource for NLP needs.

The creation of digital linguistic resources such as a lexicon is essential for ensuring that languages such as mentioned earlier remain accessible in the digital age.

This study aims to develop a structured lexicon resource for Ibaloi, designed considering compatibility with NLP processes. Such resources can support applications in machine translation, chatbots, speech recognition, educational tools, and digital preservation of indigenous knowledge.

Natural Language Processing (NLP) is a subfield of Artificial Intelligence (AI). Its concern is mainly on computational linguistics, which deals with the interaction between computers and humans, thus natural language. [3]

NLP comprises multiple processes, such as text analysis, machine translation, sentiment analysis, speech recognition, named entity recognition,

summarization, and question answering. [3] Because human language is ambiguous and context-dependent by nature, NLP is complicated and requires models that can comprehend conversational grammar, semantics, and pragmatics.

1.1 Research Objectives

While dictionaries and reference materials exist for the Ibaloi language, most are designed for human use rather than computational processing. This shortage makes it difficult to build the machine-readable word lists that are essential for strong NLP tools.

The main gap is the lack of structured, annotated, and standardized Ibaloi lexicon datasets that are suitable for computational processing.

Objectives:

1. To collect and compile existing Ibaloi vocabulary with English Translation.
2. To annotate lexical entries with linguistic features (e.g., part of speech, pronunciation, sample usage).
3. To design the dataset in a machine-readable structure for NLP applications.
4. To develop a LLM-based translator using the constructed lexicon.

1.2 Relevance of the Study

This project aims to create a structured lexicon for the Ibaloi language, contributing to the advancement of NLP. It bridges the gap between traditional lexicography and modern NLP needs.

1.3 Structure of the Paper

Chapter 1 introduces the study, including background, objectives, significance, and scope.

Chapter 2 discusses the methodology used in building the lexicon.
Chapter 3 lists the references used in the study.

1.4 Scope and Limitations

This project limits the scope of Ibaloi word extraction to words that are used in common conversations. Furthermore, since there are limited personnel in this research who are capable of verifying the output in the Ibaloi language, errors in grammatical and syntactical structures might be overlooked. For the formulation of Ibaloi word structure, second-hand resources are included

2. METHODOLOGY

The following sections describe the methods and tools used in conducting the study.
This study applies a qualitative-descriptive approach as it mainly documents and structures linguistic data for NLP.

2.1 Ibaloi Vocabulary Compilation

For collecting and compiling existing Ibaloi vocabulary, various online resources were used. The main resource was online Ibaloi dictionaries and databases, where a CSV file for the words was provided.

2.1.1 Word Collection

For a diverse collection of words, web scraping was also conducted to collect words from different resources, such as social media platforms and other online sources.

2.1.2 Lexicon Columns

Column Names	
Word	The base Ibaloi word
Pronunciation	Spelled-out pronunciation
POS	Parts of Speech
englishTranslation	English counterpart of the Ibaloi word
synonyms	Ibaloi synonyms
ibaloiSentence	Sample usage of words in a sentence.
englishSentence	English translation of the Ibaloi sentence.

2.1.1 Sentence Builder Website

A website is created in order to facilitate data collection of Ibaloi sentences as the group lacks the necessary manpower and resources to create a dataset that uses the words collected from various books. The website was developed by the group using basic web development tools such as html, css, and javascript (node). An API was created connected to the Excel sheet of collected words the group collected, that automatically gathers and sends the words to the website. After finalizing the website was then deployed in Versel.

The main goal of the website is to gather Ibaloi sentences with their English counterparts from the native speakers of the language. The website flow is as follows. The user goes to the website then the website sends a request to the api to get 5 words that are not overused in the dataset. The api then sends it as a json to the website which is then unpackaged to the front end as five words. The user then picks a main word where the sentence will revolve around and construct a sentence and provide its english translation. The user can input more than one sentence by clicking the submit another sentence button.

That is all for the users. In the backend the api receives the package from the website containing the sentences and then stores it in a sheet where it is then processed into its respective columns.

2.2 Lexical Annotation

Several attributes were added in the lexicon to illustrate the use of Ibaloi language in various contexts. To allow English speakers around the world to understand the target language, "English word" is added to hold an English translation of the desired Ibaloi word; "Ibaloi Synonym" lists out any source language words that convey a similar meaning; "Ibaloi Sentence" and "English Sentence" demonstrates the usage of the target language word and language word in a sentence, allowing learners to understand how the Ibaloi word shall be placed inside a structure defined by the grammar in the Ibaloi language, along with the English counterpart.
To ensure phonetic reliability and accessibility of Ibaloi words for non-native speakers, the researchers adapted the pronunciation respelling conventions of Google Dictionary and the Oxford English Dictionary (see Appendix B). The developed system standardizes vowel and consonant representations in simplified English-based respelling.

A sentence formed while machine translation translates a source language into a target language, however, would take time to make sense of due to the misarranged sequence that would otherwise confuse speakers of the target language. By implementing part-of-speech tagging, translated words could be rearranged to fit the design of parts of speech in the target language, facilitating understanding.

2.3 Data Processing

The extracted word lists underwent a systematic process of cleaning, annotation, and normalization to ensure quality and consistency. During the cleaning phase, duplicates were removed, spelling and orthographic inconsistencies were corrected, and all entries were aligned to a uniform format. Afterward, the words were annotated with relevant linguistic information, including part of speech, simplified pronunciation, meaning or definition, and usage examples when available, with reference to established orthographic guides to maintain accuracy. Finally , normalization was performed to harmonize variants of the same word, such as borrowed forms or dialect differences, while marking them for cross-reference to preserve both linguistic accuracy and cultural authenticity.

2.3.1 Orthographic Normalization

A case that was observed in the collected words is the orthographic variations, where the same word appears in different spellings but has the same meaning and pronunciation. The criteria that was applied in choosing the standard spelling is choosing the most frequently used spelling, a preferred spelling from the native speakers, and the ones that has the linguistically simplest form.

Alternative spellings were retained as recorded variants and cross-referenced to the standard entry. This method follows best practices in documentary linguistics and lexicographic standardization (Himmelmann, 1998; Kilgarriff, 2012; UNESCO, 2011). This approach ensured spelling consistency while preserving the natural variation present in community usage.

2.4 Application Creation

With the help of LLM APIs and an API key, developers can invoke LLMs to serve inside an application. The developers would then devise prompts, instructing it on how to interpret the electronic lexicon and come up with a reply accordingly. In this case, the prompt instructed the LLM to receive a rough translation made through machine translation that looked up every word a user inputted and converted it into the target language's equivalent word. Guidelines regarding syntax, grammar, orthography, spelling, tone, and style are then incorporated into the prompt to perform inference, curating words or sentences close to how native speakers would have said it otherwise. Outputs from the LLM is contained in a webpage. The web application framework Flask is built with several notable features in store, allowing users to gain access to an interactive electronic lexicon.

3. REFERENCES

[1] Villanueva, C. B. (2022). Language subject access to Indigenous materials: The Philippine Cordillera case. Cataloging & Classification Quarterly, 60(3–4), 297–314. https://doi.org/10.1080/01639374.2022.2075512

[2] https://www.scribd.com/document/505647074/Ibaloy-Orthography-FN

[3] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2022. Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications 82, 7 (2023).