A Comprehensive Dialect Conversion Approach from Chittagonian to Standard Bangla

Hafizur Rahman Milon, Sheikh Nasir Uddin Sabbir, Azfar Inan, Nahid Hossain
Department Of Computer Science and Engineering
United International University, Dhaka, Bangladesh
hafizurmilon.11@gmail.com, nasirsabbir07@gmail.com, azfar.inan0615@gmail.com, nahid@cse.uiu.ac.bd

***Abstract*—We present a comprehensive conversion system to convert the Chittagonian dialect to standard Bangla language. It is a text to text conversion system based on word-to-word mapping adopting a bilingual dictionary, rule-based morphological transformation on suffixes, and a supportive word suggestion module. The system tokenizes the regional input text and processes the tokens through word-to-word mapping and morphological transformation using suffix transformation rules if word-to-word mapping fails. We are also introducing an aiding tool that generates suggested words for the dialectal input. The system achieved an accuracy of 94.75% for producing standard Bangla translation from Chittagonian words. It must be noted that there is no published work on the Chittagonian dialect conversion from a computational point of view. We are the first ones to have built such a system for Chittagonian dialect to standard Bangla conversion.**

***Keywords*—Bangla, Dialect, Chittagonian, Double Metaphone.**

I. Introduction

According to many linguists and researchers, dialects are just a different form of the language, spoken with different accents and morphemes. A dialect may even have its own grammar and sentence rules. Some dialects are rich enough to be accepted as a full-fledged language. Chittagonian is one of the principal dialects of Bangla language that is spoken widely across the south-eastern region as the only means of communication. It is one of the most intricate dialects for the non-native standard Bangla speakers to understand as it is rich with its words and phrases. Activities like establishing deals and finding accommodations prove to be challenging from time to time. To cope with this, people are using English and standard Bangla more; as a result, this enriched dialect is losing its speakers day by day.

As we have mentioned earlier, no notable work has been done yet that deals with the conversion of the Chittagonian dialect. In 2017, Amrita Das presented an in-depth study on Sylheti grammar which helped us to work with Chittagonian grammar [1]. Mohammad Azizul Hoque's 2015 paper on Chittagonian language describing Chittagonian grammar, word pronunciation which helped us with our research [2]. In 2015, Arvinder Singh et al. proposed a converter for Punjabi dialects that worked using a rule-based approach and bilingual dictionary [3]. In 2014, K Marimuthu et al. provided a method to convert dialectal Tamil text to standard Tamil text using Finite State Transducers, which yielded an accuracy rate of 85% [4]. In 2012, G.H. Al-Gaphai et al. worked with 9386 words and their rule-based approach yielded an accuracy of 77.32% [5]. Hitahm Abo Bakr et al. proposed a hybrid approach for converting Egyptian colloquial to Modern Standard Arabic with an accuracy of 88% in 2008 [6]. They used tokenization and POS (Parts of Speech) tagging to improve the performance of their system. Md. Shahnur Azad Chowdhury worked on Bangla to English machine translation using POS tagging [7]. He used Tag Vectors and a set of grammar rules for the conversion process.

Our proposed system is the first that provides a comprehensive solution. We have created a bilingual dictionary as the dataset to map standard Bangla word for Chittagonian word. If the word-to-word mapping fails to give a proper translation, the system moves to suffix transformation. It splits each token into a root and a suffix and performs word-to-word mapping on the root word. We have used POS tagging to find the proper suffix that fits with the standard Bangla root word. We have also provided a word suggestion module since people might spell the same word differently. We acquired the suggestions by means of Double Metaphone Encoding [8], LCS (Longest Common Subsequence) [9] [10], and K-NN (K-Nearest Neighbors) [11]. Double Metaphone algorithm encodes the input into corresponding English letters, LCS compares Double Metaphone encodings to determine similarity and K-NN finds the closest matches to generate the suggestions.

Section II describes the proposed system and presents step by step explanation of our work along with algorithms. The experimental results and performance analysis is provided in section III, section IV concludes the paper with limitations of the system and future work.

II. Proposed Method

In this section, we have incorporated the whole process step by step in detail.

A. Dataset Collection and Corpus Study

Chittagonian dialect has a very different set of words than that of standard Bangla. The key part of the converter is the dataset. Accuracy and time complexities are immensely dependable on the dataset alone. Chittagonian dialect hardly has any resources in written format. Although it's enriched in culture and literature, it lacks written texts, especially in a

978-1-7281-7366-5/20/$31.00

digital format. We had to build the dataset from scratch. We have collected most of the data from the book [12] by Noor Mohammad Rafiq. The book has 8500 Chittagonian words along with Bangla translation and about 100 complete sentence examples. Secondary sources were websites and social media posts and comments. Finally, our dataset contains 20,101 Chittagonian root words and 5,010 complete Chittagonian sentences for rule generation and 2,230 complete Chittagonian sentences for testing the system's performance. After studying the corpus, we have noticed no significant differences between Chittagonian and standard Bangla in terms of grammatical rules. The key differences were words and suffixes and in some negative sentences. Also, we have noticed that the spelling of a particular word may differ from person to person based on their accents and preferences.

B. Translation Methodologies

We have used Tokenization, Rule-based Negation Handling, Word-to-Word Mapping, and Morphological Transformation using suffix rules. Fig 1 shows the entire process of the Translation module.

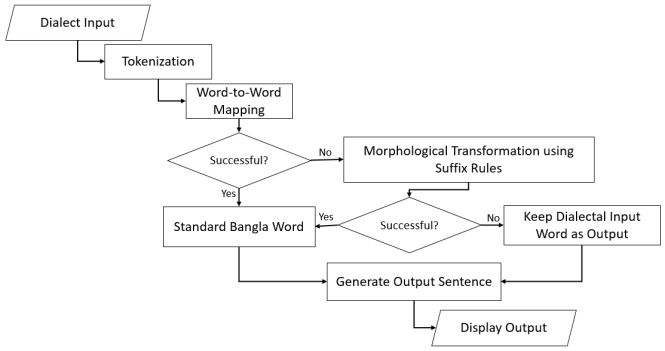

Figure 1. System Diagram of Translation module.

1) Tokenization: In tokenization, we split the input sentence into separate words or tokens using Python's basic split method and prepare those tokens for further processing. For example, the sentence 'হাই ন ডরাই' is split into 'হাই', 'ন' and 'ডরাই'.

2) Rule-based Negation Handling: In most of the Chittagonian negative sentences, the negative word 'ন' precedes the Verb (aka. ক্রিয়া) whereas it follows the Verb in standard Bangla. For example, in the sentence 'হাই ন ডরাই', the negative word 'ন' comes before the Verb while in standard Bangla sentence 'আমি ভয় পাই না', the negative word 'না' sits after the Verb. We handle these negations based on rules we have generated after analyzing Chittagonian sentences with negations.

3) Word-to-Word Mapping: This is the most important phase of the conversion process. Each input word is mapped into its corresponding standard Bangla word. It is a simple one-to-one mapping based on the input word.

Some Chittagonian words have different standard Bangla meanings. That creates a one-to-many relationship. For example, the word 'আই' means both 'আমি' and 'এসে' in different contexts. At this stage of our work, we omit processing words based on their context. Some examples of word-to-word mapping between Chittagonian and Bangla words are given in Table I.

Table I
WORD-TO-WORD MAPPING EXAMPLES.

Chittagonian Word	Standard Bangla Word
অনেরা	আপনারা
অলপল	অগোছালো
ফইয়া	ভিক্ষুক
ইতারে	তাকে
আইন্দা	আগামী

Algorithm 1: Translation

```
inputSentence ← str()
tokens ← Tokenize(inputSentence)
Handle for negative sentences if any
for each ctgWord in tokens do
    bngWord = Translate(ctgWord)
    if bngWord = "" then
        Split ctgWord into ctgStem and ctgSuffix
        bngStem = Translate(ctgStem)
        if bngStem = "" then
            bngWord = ctgWord
        else
            Get bngSuffix based on bngStem and ctgSuffix
            bngWord = bngStem + bngSuffix
        end if
        translation+ = bngWord + "space"
    end if
end for
```

4) Morphological Transformation Using Suffix Rules: If word-to-word mapping fails to generate a translation (i.e. returns empty string), the system processes the input word using a rule-based approach. The system splits the input word into stem and suffix utilizing a collection of Chittagonian suffixes and translates the stem to a standard Bangla stem using word-to-word mapping. Then the Chittagonian suffix is mapped into the corresponding standard Bangla suffix using transformation rules. Finally, the system adds the output root word and the suffix to generate the standard Bangla word.

We have noticed that the suffix transformation rules are variant depending on the last character of the standard Bangla root. Different suffixes are produced for the last character to be a vowel (স্বরবর্ণ) or a consonant (ব্যঞ্জনবর্ণ). For example, the Bangla root words 'দেশ' and 'দুনিয়া' in table II transformed the same Chittagonian suffix 'ত' differently. The rules are also different on the same suffix for different POS. For example, 'দেশত' => 'দেশে' (বিশেষ্য) and 'টাইলত' => 'কাটাত' (ক্রিয়া) with

the same suffix 'ত'. Some example rules are shown in table II.

Table II
SUFFIX RULES EXAMPLES.

Chittagonian Word	Bangla Stem	Bangla Last Character	Bangla Word
দেশত = দেশ + ত	দেশ	Consonant	দেশে = দেশ + ে
দুন্নাইত = দুন্নাই + ত	দুনিয়া	Vowel	দুনিয়ায় = দুনিয়া + য়
দুক্কে = দুক্ক + ে	দুঃখ	Consonant	দুঃখে = দুঃখ + ে
কাদনর = কাদন + র	কান্না	Vowel	কান্নার = কান্না + র
ফইরারে = ফইরা + রে	ভিক্ষুক	Consonant	ভিক্ষুককে = ভিক্ষুক + কে
পাত্তরগান = পাত্তর + গান	পাথর	Consonant	পাথরটি = পাথর + টি

C. Word Suggestion Methodologies

There are no conventional spelling rules for Chittagonian dialect. The spelling of a particular word may differ from person to person based on their accents and preferences. For example, the word 'অনেরা' from one user may be spelled differently as 'অনারা' or 'হনেরা' by another. Both of them are potentially correct meaning 'আপনারা' in standard Bangla. This could be catastrophic for the system as the system might fail to generate a correct translation. We have introduced the word suggestion module to tackle this problem.

This module checks the input words and provides the user with a collection of suggested words for each input word. Key techniques used in the word suggestion module are: Double Metaphone Encoding, LCS and K-NN. Table III shows some sample outputs of the word suggestion module.

Table III
WORD SUGGESTION EXAMPLES.

Input Word	Suggested Words
অনেরা	অনেরা, অনারা, ঐন্না
উযু	উযু, উযুউযু, উইযুই
আননি	আননি, আনটেংসি, আনযা-আনযি
ইয়ত	ইয়ত, অছিয়ত, ঐয়ত

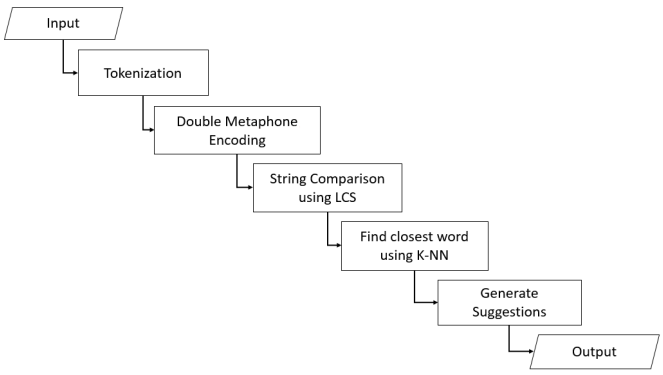

Figure 2. Word Suggestion System Diagram.

Algorithm 2: Word Suggestion

$inputSentence \leftarrow str()$
$tokens \leftarrow Tokenize(inputSentence)$
for each wordX in input **do**
  $dmX \leftarrow DMetaphone(wordX)$
  for each wordY in Chittagonian **do**
    $dmY \leftarrow DMetaphone(wordY)$
    $dmXLen = len(dmX)$
    $dmYLen = len(dmY)$
    if $dmXLen > dmYLen$ **then**
      $lcsLen = LCS(dmX, dmY)$
    else
      $lcsLen = LCS(dmY, dmX)$
    end if
    Add lcsLen to lcsLenList
  end for
  Find top 3 suggested words with highest lcsLen using K-NN
  return suggestions
end for

1) Double Metaphone Encoding: The given words in the input sentence are encoded using Double Metaphone Encoding. We have implemented Mumit Khan's Double Metaphone encoding table [14]. This process encodes the Bangla alphabets into corresponding English characters. Each Bangla character is coded with one or more English characters based on different contexts of the word. In our Double Metaphone encoding, we coded only consonant characters (ক,খ...). We didn't encode vowels (অ,আ...), since vowels do not put significant difference in pronunciation of a word [13] [14]. Some examples are provided in Table IV.

Table IV
DOUBLE METAPHONE ENCODING EXAMPLES.

Chittagonian Word	Double Metaphone Encoding
অনেরা	onera
অলপল	olpl
ফইয়া	piya
ইতারে	itare

2) Longest Common Subsequence: LCS finds all the possible subsequences in a string and also outputs the largest subsequence existing in a string. The Double Metaphone encodes of input words are used for string matching using LCS. It generates the LCS length of two words that is the length of the longest match between them. For each input word, a list of LCS lengths is generated for all the Chittagonian words. The list is then passed onto the K-NN process. Examples are provided in Table V.

3) K-Nearest Neighbors: We've used K-NN to find the words that match the highest with input word and output as suggested words. The LCS lengths work as the distance attribute here. The highest LCS length means the lowest distance and the lowest means the highest. All it does is find the words with the highest LCS lengths and suggest words

placing the closest word at the top of the suggestion list. We pass the LCS length list as shown in table V onto K-NN, then, the system processes the LCS lengths of each word and find the closest ones. Here, K is set to 3 which means it finds 3 closest suggestions.

Table V
LCS LENGTH EXAMPLES.

Input Word	Chittagonian Word	D.M. Encoding	LCS Length
অনেরা	অনেরা	onera	5
	অনারা	onara	4
	ঐন্না	oinna	3
	আনৈননা	anoinna	3
	ইতারা	itara	2

III. Experimental Results And Performance Analysis

In this section, we have demonstrated the results and performance analysis of translation and word suggestion module.

A. Translation Evaluation

We have tested our system's performance with 2,230 (11,042 words) Chittagonian sentences. We have divided these sentences into 4 smaller test sets and tested our system individually. If there is a single mistake anywhere (word mapping, suffix/prefix rule, and punctuation) in the output sentence, we have considered the whole sentence as an incorrect/erroneous conversion. The results are shown in the table below.

Table VI
EXPERIMENTAL RESULTS.

Dataset Name	Sentence Count	Error Count	Accuracy(%)
test1	557	26	95.332
test2	557	35	93.716
test3	558	29	94.803
test4	558	27	95.161

From the information available in table VI, we have calculated an average conversion accuracy rate of 94.75%.

From close observation, we have noticed that most of the errors occurred due to one-to-many mapping of a word (i.e. repetition of the same word with different meanings). For example, the word 'হক্কল' has two different meanings: 'সকল' and 'সবাই'. Among other reasons were the lack of pure suffix transformation rules and the difference between 'Sadhu (সাধু)' and 'Chalit (চলিত)' accents.

B. Word Suggestion Evaluation

The Word Suggestion module evaluation process was a bit tricky since a particular input word may have different correct suggestions based on different user input and requirements. There are no correct or incorrect results in this module. For example, the input word 'ওনেরা' (misspelled) gave us the output suggested words 'অনেরা', 'অনারা' and 'আনসাড়া'. The words 'অনেরা' and 'অনারা' both were pretty similar to the input. But the word 'আনসাড়া' seemed very dissimilar to the input. We think that the lack of dataset or words is the main reason behind the one peculiar output. A larger dataset would nudge the module towards generating more relevant suggested words.

IV. Conclusion And Future Work

We have presented a comprehensive dialect converter for the Chittagonian dialect and demonstrated different structures of Chittagonian dialect. We have built a dataset of Chittagonian words, implemented word to word mapping, morphological rules for translation, and a module for word suggestion. Our method yields an encouraging result at this stage of our work. The main limitation of our work is the size of the dataset. We are working on increasing its size to attain better usability. Learnability of the system is another big issue as we have not used any machine learning algorithms for the system to improve itself. We are working on the implementation of Neural Networks for the system to make it more robust and to increase the accuracy of the suggestion words. Authors are currently working on the implementation of an STT (Speech to Text) and a TTS (Text to Speech) for our system.

References

[1] Amrita Das, "A Comparative Study of Bangla and Sylheti Grammar," PP. 389, Università degli Studi di Napoli Federico II, 2017.

[2] Muhammad Azizul Hoque, "Chittagonian Variety: Dialect, Language, or Semi-Language?,"CRP, International Islamic University Chittagong, Bangladesh, 2015.

[3] Arvinder Singh and Parminder Singh, "Punjabi dialects conversion system for Malwai and Doabi dialects," Vol.8, PP.1–6, Indian Journal of Science and Technology, 2015.

[4] K Marimuthu and Sobha Lalitha Devi, "Automatic conversion of dialectal Tamil text to standard written Tamil text using FSTs," PP. 37–45, Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, 2014.

[5] GH Al-Gaphari and M Al-Yadoumi, "A method to convert Sana'ani accent to Modern Standard Arabic," Vol.8, PP. 39–49, International Journal of Information Science and Management (IJISM), 2012.

[6] Hitham Abo Bakr, Khaled Shaalan and Ibrahim Ziedan, "A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic,"The 6th international conference on informatics and systems,Cairo university, 2008.

[7] Md Shahnur Azad Chowdhury, "Developing a Bangla to English Machine Translation System Using Parts Of Speech Tagging," Vol.1, PP. 113-119, Journal of Modern Science and Technology, 2013.

[8] Naushad UzZaman and Muhit Khan, "A double metaphone encoding for Bangla and its application in spelling checker," PP. 705-710, International Conference on Natural Language Processing and Knowledge Engineering, 2005.

[9] Lasse Bergroth, Harri Hakonen and Timo Raita, "A survey of longest common subsequence algorithms," PP. 39-48, Proceedings Seventh International Symposium on String Processing and Information Retrieval, SPIRE 2000.

[10] Deena Nath, Jitendra Kurmi and Vipin Rawat, "A Survey on Longest Common Subsequence," vol. 6, International Journal for Research in Applied Science & Engineering Technology (IJRASET), 2018.

[11] Sayali D. Jadhav and HP Channe,"Comparative study of K-NN, naive Bayes and decision tree classification techniques," Vol. 5, PP. 1842–1845,International Journal of Science and Research (IJSR), 2016.

[12] Noor Muhammad Rafiq, "Chottogramer Ancholik Bhashar OVIDHAN,"ISBN:9789849107521, 2nd edition,2017.

[13] Min-Siong Liang, Ren-Yuan Lyu and Yuang-Chin Chiang, "Phonetic transcription using speech recognition technique considering variations in pronunciation," Vol. 4, PP. IV–109, 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, 2007.

[14] Naushad UzZaman and Mumit Khan. "A Double Metaphone Encoding for Approximate Name Searching and Matching in Bangla." Computational Intelligence (2005).