

Inclusive bidirectional conversion system between Chittagonian and standard Bangla

Nahid Hossain¹, Hafizur Rahman Milon², Sheikh Nasir Uddin Sabbir¹, Azfar Inan^{1,3}

¹Department of Computer Science & Engineering, United International University, Dhaka, Bangladesh

²Junior Software Engineer, Barikoi Technologies Limited, Dhaka, Bangladesh

³Software Engineer, Celloscope Limited, Dhaka, Bangladesh

Article Info

Article history:

Received Oct 14, 2021

Revised Dec 10, 2021

Accepted Jan 11, 2022

Keywords:

Bangla

Bidirectional converter

Chittagonian dialect

Dialect converter

ABSTRACT

In recent years, the Bangla language has come out as a very prominent figure in the world of natural language processing. Many researchers have produced exemplary works on the language, but there has not been any notable work on its highly enriched dialects due to the lack of resources and the diversity and complexity in its grammatical structure. This paper has suggested a bidirectional conversion system for one of the most widely used dialects of Bangla; the Chittagonian dialect. Our method employs a bidirectional lexicon that uses binary search, word-to-word mapping, and morphological transformations. The system has achieved an accuracy rate of 95.86% for Chittagonian to standard Bangla and 93.89% in the case of standard Bangla to Chittagonian. We have also provided an efficient word suggestion module, and it has yielded satisfactory results.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Nahid Hossain

Department of Computer Science and Engineering, United International University

Dhaka, Bangladesh

Email: nahid@cse.uui.ac.bd

1. INTRODUCTION

Chittagonian is one of the major dialects of the Bangla language, broadly spoken as the sole language in Bangladesh's south-eastern region. It is the most difficult dialect for non-native standard Bangla speakers to understand due to its abundance of words and phrases. Ventures such as negotiating deals and obtaining lodging can be challenging at times. Hence, this enriched and historical dialect of Bangla language is losing speakers regularly. So far, the conversion of the Chittagonian dialect has received relatively little attention. In this paper, we have suggested a bidirectional conversion system for one of the most widely used dialects of Bangla; the Chittagonian dialect. We have built a bidirectional dictionary to map the standard Bangla words with Chittagonian words and vice-versa. If word-to-word mapping provides an invalid translation, the system moves to morphological transformation. Each token is divided into a root and a suffix, and the root term is mapped accordingly. We have developed suitable rules to find the correct suffix for the standard Bangla root words. Since people often spell the same word in different ways, we have also introduced a word suggestion module. We have obtained the list of suggestions using double metaphone encoding [1]-[3] since the encoding can produce multiple encodings for the same word with different pronunciations, Jaro-Winkler similarity algorithm [4], [5] since the algorithm gives higher accuracy in multi-byte Bangla characters than other similar algorithms and (K-nearest neighbors) K-NN since it gives higher accuracy in suggesting closest Bangla words [6]. The double metaphone encoding method has been used to encode the input into corresponding English letters, the Jaro-Winkler similarity algorithm compares the encodings to determine similarities, and K-NN generates suggestions based on the nearest matches.

Chittagonian is a vastly used but highly regional dialect. As a result, it is considerably hard for non-native speakers to understand which, creates a hindrance in communication. In the long term, this might lead to the dialect's extinction. A standard language does streamline the official and economic procedures but, it can also have a long-term impact on the language treasury. These concerns have motivated us to develop a proper bidirectional dialect converter for the Chittagonian dialect which, can help mitigate these issues. The contributions of the proposed work are as: i) a digital and structured Chittagonian dataset; ii) a collection of grammatical rules of the dialect; iii) an efficient bidirectional translator with word suggestions module. Section 2 provides a brief overview of various works on conversion approaches, section 3 explains the proposed system and provides an extensive explanation of our work, including the algorithms. Section 4 contains the study findings and performance analysis, while section 5 ends the paper with its shortcomings and future work.

As the work done on a conversion system for any dialect for the Bangla language is minimum, we have studied the approaches taken by researchers for other languages. The study performed by Arpita Goswami on the Sylheti grammar in 2021 has helped us understand the Chittagonian grammar [7]. Due to the limited work available on Bangla language conversion, we have studied the procedures taken by researchers of other major languages. Such as the converter suggested by Singh and Singh [8] in 2015 to convert the Punjabi dialect by adopting a rule-based approach and a bilingual dictionary and the recommended system by K and Devi [9] in 2014 for a converter of indigenous Tamil text to standard Tamil text by employing finite state transducers, which produced an efficiency of 85%. Some significant works have also been performed on the Arabic language. Al-Gaphai and Yadoumi [10] suggested a rule-based approach in their 2012 paper that produced a veracity of 77.32% while working with 9386 words. In 2008, Bakr and Ziedan [11] suggested a hybrid method to convert to Modern Standard Arabic from Egyptian colloquial. They adopted tokenization and parts of speech (POS) tagging to enhance the performance of the model. Their suggested method achieved a precision of 88%. Chowdhury [12] proposed a POS tagging method for Bangla to English machine translation. For the conversion system, he used tag vectors and a collection of grammar rules.

2. RESEARCH METHOD

In this section, we have illustrated the outline and development methodology of our proposed system.

2.1. Dataset collection and corpus study

Although being enriched in vocabulary, ethnic music, and folklore, literary works of Chittagonian dialect are nominal in traditional written format, especially in digital format. For our system, we needed a bilingual dictionary dataset. Digital records being insufficient, we built our glossary with the help of the book [13]. The book has remained our primary source of data containing 1000 complete sentences examples and 8,500 Chittagonian words along with their corresponding Bangla translations. The secondary sources of the Chittagonian words and equivalent Bangla words are renowned Bangladeshi national newspapers, social media posts, and voluntary comments.

The dictionary dataset has three columns: Chittagonian word, standard Bangla word, and POS collected from different sources and some are manually annotated by the authors). To achieve better performance, we have divided the entire dataset into sub-datasets based on the first character of each Chittagonian word rather than keeping the glossary as a single entity. Thus allowing the system to read the dataset like a physical dictionary as a human would. The system uses the decimal code which is a simple integer ASCII value of the first Bangla Unicode character to hash/index the sub-dataset. A short snippet of the sample sub-datasets is displayed in Table 1. It shows that the sub-dataset of the character 'ক' is indexed at 2453 which is the character decimal code (ASCII value) of 'ক'.

Table 1. Sample of dictionary dataset

Index: 2453		
Chittagonian	Standard Bangla	POS
কইলা	কয়লা	বিশেষ
কালো	কালো	বিশেষণ
কেঅন	কেমন	ক্রিয়াবাচক বিশেষণ

Each sub-dataset is then sorted alphabetically in ascending order based on the Chittagonian column. The sorting is mandatory for binary search [14] to work precisely. The precision and the swiftness of the system rely heavily on the dataset. We have further improved the dataset by removing extra whitespaces, recurring entries, and anomalies and making the dataset as adequate as possible. Currently, the dataset includes 20,101 unique entries for word-to-word mapping, and for rule generation 5,010 complete Chittagonian sentences. To

test the system's accuracy and performance, the dataset has 2,230 independent Chittagonian sentences. For the standard Bangla to Chittagonian translation module, the same dataset has been used with minor modifications. In this case, the dataset is split and indexed depending on the first character of the standard Bangla column entries. Each sub-dataset is sorted by the standard Bangla column.

2.2. Translation methodologies

Our system uses the word-to-word mapping method as the primary source for transforming the provided input. In this study, we have presented a tangible method to convert the Chittagonian dialect to standard Bangla. We have also implemented the reverse conversion of standard Bangla to Chittagonian. Our system uses tokenization, word-to-word mapping, and a rule-based approach for both conversion purposes. We will be discussing the process of conversion from Chittagonian to standard Bangla in detail, followed by the reverse translation process.

2.2.1. Tokenization

Tokenization involves breaking down the input paragraph into sentences first. Then, they are broken into individual words or tokens. These tokens are used further down the process. For example, the sentence 'দুন্নাইত তার কেঅয় ন আছিল' is split into 'দুন্নাইত', 'তার', 'কেঅয়', 'ন', 'আছিল'. These tokens are then passed through to the subsequent phases of the system.

2.2.2. Rule-based negation handling

From our observation, we have found out in most cases for negative Chittagonian sentences, the negative word 'ন' precedes the verb (ক্রিয়া) whereas it follows the verb in standard Bangla. From the aforementioned example 'দুন্নাইত তার কেঅয় ন আছিল', we can see that the negative word 'ন' comes before the verb while in standard Bangla sentence 'দুন্নিয়াতে তার কেহই ছিল না', the negative word 'না' sits after the verb, like so in Figure 1. We have established this as a general rule of thumb and reformed the negative sentences by swapping the negative word with the verb. Thus, the system can provide an accurate and meaningful translation for negative sentences.

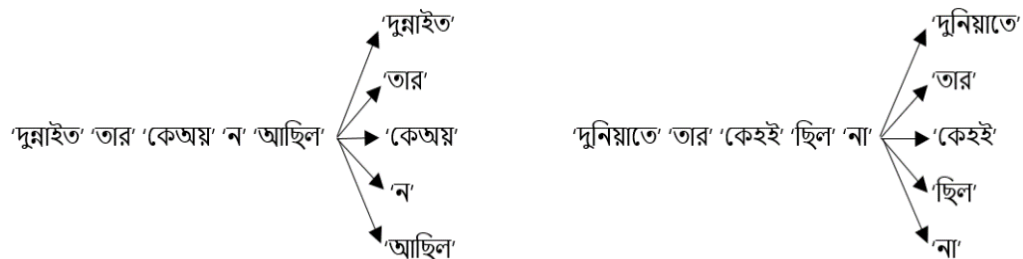


Figure 1. Negation handling

2.2.3. Word-to-word mapping

It is the most crucial step of the conversion procedure [15]. Every token is mapped to its standard Bangla equivalent using the glossary. It is a basic one-to-one mapping based on the input word. Let us take the previously mentioned sample sentence 'দুন্নাইত তার কেঅয় ন আছিল' as an example. After tokenization and handling the negative part of the sentence, the system searches for each token. The first token 'দুন্নাইত' is looked up in the lexicon. We have used a binary search method to traverse the complete lexicon before. To search faster, we divided the complete dataset into individual characters, as mentioned earlier. The system takes the tokens' first letter as a key to access the dataset. In this case, the first letter 'দ' determines the sub-dataset to look into having its character decimal code '2470' as the index. Each sub-dataset has been sorted alphabetically in ascending order. Then the system searches for the exact word in this isolated sub-dataset using binary search. Since this token does not exist in the glossary, the system can not generate the standard Bangla term. The system then feeds the token into the morphological transformation module to translate using Suffix Transformation rules. The process is further has been elaborated in the following section. Similarly, the system searches the remaining tokens in the corresponding sub-datasets and maps them into the standard Bangla entry if found. For example, the token 'আছিল' is looked up into the sub-dataset with the index '2438' (Character Decimal Code of 'আ' and produces the standard Bangla word 'ছিল' as it exists in the glossary. Table 2 provides some samples of the word-to-word mapping.

Table 2. Word-to-word mapping examples

Chittagonian word	Standard Bangla word
আছিল	ছিল
বেড়া	পুরুষ
লেছা	আঠা

2.2.4. Morphological transformation using suffix transformation rules

If word-to-word mapping produces an invalid translation as a null string, the system employs a rule-based approach to process the input word. The method divides the input word into stem and suffix using a list of Chittagonian suffixes and then uses word-to-word mapping to translate the stem word to the standard Bangla stem. The suffixes are processed further with the help of transformation rules, and equivalent standard Bangla suffixes are generated. From the sample sentence 'দুন্নাইত তার কেঅয় ন আছিল', we have seen that the word 'দুন্নাইত' has not been found in the dataset with the suffix 'ত' in it. The system splits the word 'দুন্নাইত' into 'দুন্নাই+ত'. Word-to-word mapping is applied on the stem 'দুন্নাই' which provides the equivalent standard Bangla stem 'দুনিয়া'. Afterwards, the suffix 'ত' is converted into the standard Bangla suffix 'তে' using derived transformation rules. Finally, the standard Bangla stem and suffix are concatenated to generate the translated word 'দুনিয়াতে'. If the system does not produce a translation, the original input word is kept as the output word. This small step can improve the overall accuracy of the system. Some words have the same spelling and pronunciation in both dialects. For example, the word 'নাম' means the same in Chittagonian and standard Bangla. From an extensive study on both dialects, we have discovered that the suffix transformation rules differ from time to time based on the concluding character of the standard Bangla root word. Different suffixes are formed depending on whether the concluding character is a vowel (স্বরবর্ণ) or a consonant (ব্যঞ্জনবর্ণ). For instance, the root words 'আগুন' and 'দরজা' in Table 3 transformed the same Chittagonian suffix 'ত' in a distinct way. For different POS on the same suffix, the rules are different. For example, 'দেশত' => 'দেশে(বিশেষ্য)' and 'টাইলত' => 'কাটাত(ক্রিয়া)' with the same suffix 'ত'. We have used our own Chittagonian POS tagger to determine the POS of each word and apply rules accordingly. Some example rules are shown in Table 3.

Table 3. Suffix rules examples

Chittagonian word	Bangla stem	Bangla last character	Bangla word formation
হানশর = হানশ + র	বাতি	Vowel	বাতির=বাতি + র
আদরগান = আদর + গান	মুখ	Consonant	মুখটি=মুখ + টি
দোয়ারত = দোয়ার + ত	দরজা	Vowel	দরজায়=দরজা + য়
হৌরে = হৌর + ে	শ্বশুর	Consonant	শ্বশুরে=শ্বশুর + ে

2.2.5. Generate standard bangla output

This is the final stage of the translation process. In this stage, the translated standard Bangla tokens are merged to generate the corresponding sentences and paragraphs. The output words are received from three sources: word-to-word mapping, morphological transformation, and the original input word kept as it is. The translated tokens are added together to form the output sentences and the output paragraph by joining the output sentences. Since the dialects do not differ in punctuations, they are kept the same. Figure 2 provides a simple graphical representation of our desired output.

2.2.6. Standard bangla to Chittagonian

After reanalyzing our proposed system, we have realized it would be considerably easier to translate from standard Bangla to Chittagonian. The method remains the same with a few modifications. Firstly, the deconstruction of the input remains unchanged. The system generates tokens from the input. If the input is negative, the system resolves it using the negation handling rule. Secondly, the mapping of the tokens is also similar. The system uses the same glossary for this purpose. The difference is that this time we have split the dataset according to the first character of the standard Bangla entries. The system searches the input token in the standard Bangla column and extracts the equivalent Chittagonian entry.

If the system does not find a suitable Chittagonian replacement for the standard Bangla token in the glossary, it moves on to the morphological transformation module. In the case of standard Bangla to Chittagonian, the system determines whether the concluding character of the Chittagonian stem is a constant (ব্যঞ্জনবর্ণ) or a vowel (স্বরবর্ণ) and assigns an appropriate suffix to the stem. For example, if the input sentence contains a word such as 'দুনিয়াতে' the system maps it with word-to-word mapping first. If not found, it follows the rule and splits the word into 'দুনিয়া + তে'. Then the system looks for the corresponding Chittagonian word for the stem 'দুনিয়া' in the dataset, which is 'দুন্নাই'. After locating it, the system determines the Chittagonian words' concluding letter and assigns the appropriate suffix accordingly. For the Chittagonian word 'দুন্নাই', the

appropriate suffix would be 'ত'. Thus it produces the desired translated token 'দুনিয়াত'. Finally, the translated tokens are assembled to generate the Chittagonian output. In case both the mapping and the transformation methods produce an invalid translation, the system keeps the initial term as the output. Table 4 shows some example outputs of the bidirectional translation module.

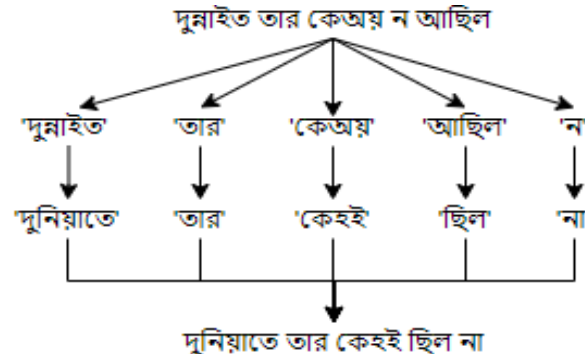


Figure 2. Standard Bangla output

Table 4. Input/output sample

Chittagonian input	Standard bangla output/input	Chittagonian output
আতেত গঅম ন লাগের।	আমার ভাল লাগছে না।	আআর গঅম ন লাগের।
অনেরা কেন আছো?	আপনারা কেমন আছেন?	অনেরা কিমান আছো?
তোয়ার নাম কি?	তোমার নাম কি?	তোয়ার নাম কি?

2.3. Word suggestion methodologies

From our study, we have found that the Chittagonian dialect might have several spellings for one word. This creates an ambiguity of choice because of varying accents and preferences from person to person. For example, the word 'অনেরা(Anera)' from one user may be spelt differently as 'অনারা(Anara)' or 'হনেরা(hanera)' by another. Both of them have a plausible correct meaning 'আপনারা(Apnara)' in standard Bangla. This might prove to be disastrous for the system since it could struggle to produce a proper translation. To address this problem, we have implemented the word suggestion module. This module validates the input words and returns a list of words as a suggestion for each input word. The following methods mentioned in subsection 3.3.1 to 3.3.4 are used for a better word suggestion. Table 5 shows some sample outputs of the word suggestion module.

2.3.1. Double metaphone encoding

The double metaphone encoding table [16], [17] has been implemented as the encoder. Based on the probable phonetics of the expression, each Bangla letter has been encoded with one or more English letters. In the proposed encoder, consonants (ক,খ...) get primarily encoded value and vowels (অ,আ...) get secondary encoded value to represent the Bangla characters since double metaphone algorithm generates different encodings for the same Bangla words and giving high priority to the Bangla consonants will avoid diverse accent and pronunciation constraints amongst different dialect speakers. The vowels determine the pronunciation, which varies from person to person resulting in various spellings [17], [18]. This proves problematic for the system to provide an accurate suggestion. Table 6 provides examples for the encoded words.

Table 5. Suggestion sample

Input word	Suggestions
মুনি	মনি মনিষ মুন্শি

Table 6. Double metaphone encoding samples

Chittagonian word	Double metaphone encoding
হতুন	httun
মুনি	muni
খ্যালা	keala

2.3.2. Jaro-Winkler similarity

We have studied various string matching algorithms such as LCS [19], [20], Rabin-Karp [21], [22], Jaro-Winkler similarity. to implement for our system. Among these, the Jaro-Winkler similarity proved to be

the most efficient method. The Jaro-Winkler similarity is used to measure the similarity between two strings [4]. This is an improvement of the Jaro similarity algorithm. The mathematical definition of the algorithm is as:

$$S_w = S_j + P * L * (1 - S_j) \quad (1)$$

Where S_j is Jaro similarity, S_w is Jaro-Winkler similarity, L is the length of the matching prefix before we have found inequality between the two strings up to a maximum of 4 characters, and P is a scaling factor that has a default value of 0.1. Jaro-Winkler similarity algorithm has proven to be a rather efficient solution for producing the suggestions than others. The accuracy of the suggested words has also increased. The prefix scaling factor has proven to be a significant influence on this. The prefix scaling factor 'P' provides a more precise suggestion when the strings share a standard prefix up to a maximum length of 'L'.

2.3.3. Generate similarity matrix

The system uses the aforementioned double metaphone encoding, Jaro-Winkler similarity, and a collection of sub-dataset to generate a similarity matrix. This matrix is passed on to the next phase for desired outputs. The system reads the input Chittagonian word and encodes it using double metaphone encoding. Again like the mapping method in the translation section, it accesses the sub-datasets. Nevertheless, unlike the translation module, it accesses a collection of sub-datasets to find similar terms rather than just the sub-dataset of the starting character. This collection includes all of the input word's characters as well as their phonetically identical characters. Let's take the Chittagonian word 'হতুন(hattun)' as an example. After studying the language, we have found the characters 'অ(a)' and 'ও(o)' are very similar to the starting character of the input word 'হ(ha)'. These characters along with the remaining characters 'ত(ta)', 'ন(na)' and their phonetically similar characters make up the collection of sub-datasets to be searched. Now that the system has decided on the starting letters of probable similar words of 'হতুন', every Chittagonian word of this collection is encoded. The system runs a comparison between these encoded words and the input. This comparison provides a similarity score S_w , ranging from 0 to 1. Here 0 means no similarity between the words, and 1 means they are identical. These scores and their corresponding words populate the similarity matrix. It is sorted in descending order based on the similarity score. Finally, the similarity matrix is passed on to the K-NN phase.

2.3.4. K-nearest neighbors

The system uses K-NN to identify the terms that have the highest correlation with the input. The generated S_w scores for similar words are stored in descending order. Hence, putting the most similar words on top of the list. The system then suggests the top 'K' number of words (in this case, the value of K is 3). K-NN is applied between Jaro Winkler similarity score against all entries. As a result, for an incorrect word 'মুনি', we get K=3 suggestions i.e. 'মুনিষ(munis)', 'মুন্শি(munsi)', and 'মুন্(mun)'.

3. RESULTS AND DISCUSSION

This section discusses the accuracy and performance evaluation process and results of the bidirectional translator and word suggestion. The evaluation process and results are elaborated in segments.

3.1. Bidirectional translator

The translation module is the key feature of the proposed system. Hence, the overall quality of the system relies on the accuracy and performance of this module. A small collection of sample input and output sentences is provided below in Table 4 for reference. The same sentences are used for both of the translation systems for comparison.

The accuracy of the module has been measured using bilingual evaluation understudy (BLEU) [23] score. It is a technique to compare machine-translated sentences to a collection of reference sentences. We have used sentence-level BLEU score evaluation for this system. The sentence-level BLEU score takes a machine-translated sentence in the target language and compares it against one or more equivalent sentences in that target language. For example, the system generated candidate standard Bangla sentence 'ইহারা খেলা খেলছে' is compared with the following reference sentences.

In Table 7, the sentences are equivalent to each other in words and meanings. The BLEU score of this particular example is measured at 0.67 using a 1 gram probability weight [24]. 1 gram BLEU score is 1 when at least one reference sentence equals the candidate sentence in tokens (Table 4 where output sentences are already meaningful candidates.) and 0 when no candidate sentence token is found in the reference sentence tokens. For our primary goal of Chittagonian to standard Bangla translation process, we measured the BLEU score with a list of 2,230 machine-translated standard Bangla sentences. The score has been an average of

0.958625(six-digit precision). That means the Chittagonian to standard Bangla translation system has an estimated 95.86% average accuracy rate. We have investigated the reasons for the candidate sentences with lower BLEU scores. The key reason we have found is the one-to-many relationship between the Chittagonian and standard Bangla words for some of the entries in the dataset. Some Chittagoian words have multiple entries with different standard Bangla meanings or different words with the same meanings in the dataset. For example, the Chittagoian word 'সেনা' has multiple standard Bangla mappings such as 'শোক' and 'আঘাত'. Some other reasons found were the lack of sufficient suffix transformation rules and the difference between 'Sadhu (সাধু)' and 'Chalit (চলিত)' accents.

Table 7. Sample reference sentences

Candidate	References
ইহারা খেলা খেলছে	তারা খেলা খেলছে
	তাহারা খেলা খেলছে
	তারা খেলছে
	তাহারা খেলছে
	তাহারা একটি খেলা খেলছে

The standard Bangla to Chittagonian translation system has been evaluated in the same process. The same collection of sentences in machine-translated Chittagonian form has been used to measure the BLEU score. This time, the BLEU score has been 0.938939 (six-digit precision) on average, which is estimated to have an average of 93.89% accuracy rate for the standard Bangla to the Chittagonian translation system. We have also extensively tested the machine-translated standard Bangla sentences using a third-party standard Bangla to English translator tool to determine the quality of the output. We have used the Google Translate API [25] for this purpose.

3.2. Word suggestion module

Table 5 shows some examples of suggestion words for some Chittagonian input words. Finding the accuracy evaluation technique of the word suggestion module for the system proved to be troublesome. Several factors came into place to justify the purity and precision of the output words, such as character sequence, different spelling styles, pronunciation, and parts of speech in some minor cases. We have designed the accuracy measurement technique combining these criteria.

Firstly, we have weighed up the character sequence and spelling similarities between the words in comparison. Secondly, the pronunciation of the words has been matched up based on double metaphone encoding. Finally, the derived output has been evaluated via user feedback to determine the accuracy of the word suggestion module. For the word suggestion module, 28 volunteers participated in the online rigorous feedback session. As it was clear after a thorough investigation that no suggested words could be eliminated from the race with a concrete set of rules, the user evaluation played a vital role in this. For example, both 'দিলে' and 'দিলুত' could be considered correct for the input word 'দিল'. The user evaluation has been conducted keeping the aforementioned factors in mind. The comparison showed a highly satisfactory accuracy and precision for the given input word. In most cases, the first three suggestions passed the test.

4. CONCLUSION

We have presented a bidirectional conversion system between the Chittagonian dialect and the standard Bangla language. The system uses sorted sub-datasets and binary search for better performance in translation. Currently, the system translates from Chittagonian to standard Bangla and vice versa. The word suggestion system employs sub-datasets to reduce time complexity, Double Metaphone encoding to process similar-sounding characters and words, and the Jaro-Winkler similarity algorithm to generate suggested Chittagonian words for the input word. Despite having significant efficiency, the system has limitations in datasets and suffix transformation rules. A larger dataset and rule-set could be vital for further improvements.





Additionally, we are continuously working on introducing deep learning techniques to formulate a generic dialect translation system that could feed on any dialectal dataset of the Bangla language and still be able to convert it into standard Bangla form. On top of that, the authors are working on developing a speech-to-text and a text-to-speech systems for the bidirectional conversion system.

REFERENCES




- [1] N. UzZaman and M. Khan, "A Double Metaphone encoding for Bangla and its application in spelling checker," *International Conference on Natural Language Processing and Knowledge Engineering*, 2005, pp. 705-710, doi: 10.1109/NLPKE.2005.1598827.
- [2] A. K. Mandal, Md. D. Hossain and Md. Nadim, "Developing an efficient search suggestion generator, ignoring spelling error for high speed data retrieval using Double Metaphone Algorithm," *13th International Conference on Computer and Information Technology (ICCIT)*, 2010, pp. 317-320, doi: 10.1109/ICCITECHN.2010.5723876.
- [3] V. S. Vykhovanets, J. Du, and S. Sakulin, "An overview of phonetic encoding algorithms," *Automation and Remote Control*, vol. 81, no. 10, pp. 1896–1910, 2020.
- [4] B. Leonardo and S. Hansun, "Text documents plagiarism detection using rabin-karp and jaro-winkler distance algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 5, no. 2, pp. 462-471, 2017, doi: 10.11591/ijeecs.v5.i2.pp462-471.
- [5] I. Ahamed, M. Jahan, Z. Tasnim, T. Karim, S. Reza, and D. Hossain, "Spell corrector for bangla language using norvig's algorithm and jaro-winkler distance," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 1997-2005, 2021, doi: 10.11591/eei.v10i4.2410.
- [6] S. D. Jadhav and H. Channe, "Comparative study of k-nn, naive bayes and decision tree classification techniques," *International Journal of Science and Research (IJSR)*, vol. 5, pp. 1842-1845, 2016.
- [7] A. Goswami, "Marked geminates as evidence of sonorants in sylheti bangla: An optimality account," *Acta Linguistica Asiatica*, vol. 11, no. 1, pp. 99-112, 2021, doi: 10.4312/ala.11.1.99-112.
- [8] A. Singh and P. Singh, "Punjabi dialects conversion system for Malwai and Doabi dialects," *Indian Journal of Science and Technology*, vol. 8, no. 27, pp. 1-6, 2015, doi: 10.17485/ijst/2015/v8i27/81667.
- [9] M. K and S. Lalitha Devi, "Automatic conversion of dialectal Tamil text to standard written Tamil text using FSTs," *Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM. Baltimore, Maryland: Association for Computational Linguistics*, 2014, pp. 37-45. [Online]. Available: <https://aclanthology.org/W14-2805>.
- [10] G. Al-Gaphari and M. Al-Yadoumi, "A method to convert sana'ani accent to modern standard arabic," *International Journal of Information Science and Management (IJISM)*, vol. 8, pp. 39-49, 2012.
- [11] K. S. H. A. Bakr and I. Ziedan, "A hybrid approach for converting written egyptian colloquial dialect into diacritized arabic," *The 6th international conference on informatics and systems, Cairo university*, 2008.
- [12] M. S. A. Chowdhury, "Developing a bangla to english machine translation system using parts of speech tagging," *Journal of Modern Science and Technology*, vol. 1, pp. 113-119, 2013.
- [13] N. M. Rafiq, "Chottogramer Ancholik Bhashar Ovidhan, 2nd ed.," Chottogram, Balaka Publication, 2017, pp. 1-280. Accessed on: Jan 10, 2022. [Online]. Available: <https://www.rokomari.com/book/181857/chottogramer-ancholik-vashar-ovidhan>
- [14] A. Mehmood, "Ash search: Binary search optimization," *International Journal of Computer Applications*, vol. 178, no. 15, pp. 10-17, 2019.
- [15] N. Gordeev, B. Kunyavskii, and P. Eugene, "Word maps, word maps with constants and representation varieties of one-relator groups," *Journal of Algebra*, vol. 500, pp. 390-424, 2018, doi: 10.1016/j.jalgebra.2017.03.016.
- [16] N. Hossain, S. Islam and M. N. Huda, "Development of Bangla Spell and Grammar Checkers: Resource Creation and Evaluation," *IEEE Access*, vol. 9, pp. 141079-141097, 2021, doi: 10.1109/ACCESS.2021.3119627.
- [17] N. Uzzaman and M. Khan, "A double metaphone encoding for approximate name searching and matching in bangla," *Computational Intelligence*, 2005.
- [18] H. R. Milton, S. N. U. Sabbir, A. Inan and N. Hossain, "A Comprehensive Dialect Conversion Approach from Chittagonian to Standard Bangla," *IEEE Region 10 Symposium (TENSYP)*, 2020, pp. 214-217, doi: 10.1109/TENSYP50017.2020.9230714.
- [19] M. Kiyomi, T. Horiyama, and Y. Otachi, "Longest common subsequence in sublinear space," *Information Processing Letters*, vol. 168, p. 106084, 2020, doi: 10.1016/j.ipl.2020.106084.
- [20] L. Bergroth, H. Hakonen and T. Raita, "A survey of longest common subsequence algorithms," *Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000*, 2000, pp. 39-48, doi: 10.1109/SPIRE.2000.878178.
- [21] A. H. Lubis, A. Ikhwan, and P. L. E. Kan, "Combination of levenshtein distance and rabin-karp to improve the accuracy of document equivalence level," *International Journal of Engineering & Technology*, vol. 7, no. 2.27, pp. 17-21, 2018.
- [22] M. Yulianto and N. Nurhasanah, "The hybrid of jaro-winkler and rabin-karp algorithm in detecting in- donesian text similarity," *Jurnal Online Informatika*, vol. 6, no. 1, pp. 88-95, 2021. 10.15575/join.v6i1.640.
- [23] S. Chauhan, P. Daniel, A. Mishra, and A. Kumar, "Adableu: A modified bleu score for morphologically rich languages," *IETE Journal of Research*, pp. 1–12, 2021, doi: 10.1080/03772063.2021.1962745.
- [24] T. Islam, A. I. Prince, M. M. Z. Khan, M. I. Jabiullah, and M. T. Habib, "An in-depth exploration of bangla blog post classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 2, pp. 742-749, 2021, doi: 10.11591/eei.v10i2.2873.
- [25] V. Kulikov and G. Yerkebulan, "About using google custom search and google translate api in detection of cross-language plagiate," *Bulletin of Almaty University of Energy and Communications*, vol. 4, pp. 109-116, 2019.

BIOGRAPHIES OF AUTHORS






Nahid Hossain     is an Assistant Professor of the Computer Science and Engineering (CSE) department of United International University (UIU), Bangladesh. Before joining UIU, He worked as a Software Engineer at eGeneration Ltd, Bangladesh. He completed his Master's degree from UIU and Bachelor's degree from Frederick University, Cyprus and UIU jointly. He has got several national and international scholarships and awards including scholarship from European Union and Gold Medal from Education Minister of Bangladesh. His research interests include natural language processing, data mining, big data, and machine learning and AI. He can be contacted at email: nahid@cse.uiu.ac.bd or www.nahid.org.






Md Hafizur Rahman Milon    is working as a Junior Software Engineer at Barikoi Technologies Limited. For over a year, he has been devoted to producing quality products related to maps, routing, and geospatial data visualization. He graduated from United International University with a Bachelor's degree in Computer Science and Engineering. His research interests include natural language processing, machine learning & AI, and map engineering. He can be contacted at email: hafizurmilon.11@gmail.com.



Sheikh Nasir Uddin Sabbir    is working as a Research Assistant in United International University, Bangladesh. He has completed his Bachelor's degree from United International University in Computer Science and Engineering. His research interests include natural language processing, machine learning, and AI. He can be contacted at email: nasirsabbir07@gmail.com.



Azfar Inan    is a Software Engineer working in Mobile App development in Celloscope Limited, Bangladesh. Before joining Celloscope, He was an Intern in Competitive programming learning department at Samsung R&D Bangladesh Limited. He completed his Bachelor's degree from UIU. His research interests include natural language processing, image processing, machine learning and AI. He can be contacted at email: azfar.inan0615@gmail.com.