An Emperical Framework of Idioms Translator From Bengali to English: Rule Based Approach

Preprint · April 2020

DOI: 10.13140/RG.2.2.30728.98568

Ayesha Khatun
Chittagong University of Engineering & Technology
20 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Md Gulzar Hussain
Green University of Bangladesh
19 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Md. Jahidul Islam
Green University of Bangladesh
24 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Sumaiya Kabir
Green University of Bangladesh
13 PUBLICATIONS **11** CITATIONS
SEE PROFILE

An Emperical Framework of Idioms Translator From Bengali to English: Rule Based Approach

Ayesha Khatun*, Md Gulzar Hussain†, Md Jahidul Islam‡, Sumaiya Kabir§, Md Mahin¶
Department of Computer Science & Engineering,
Green University of Bangladesh, Dhaka, Bangladesh.
ayeshankhatun@gmail.com*, gulzar.ace@gmail.com†, jahidul.jnucse@gmail.com‡,
summa.cse@gmail.com§, mahin@cse.green.edu.bd¶

Abstract—**Idioms are taking a vital part in effective communication as well as a crucial part of cultural inheritance. It represents the group of words together have the meaning which is different from an individual word meaning, for this metaphorical behavior idioms arise difficulties in the general machine translation system. In this paper, we have proposed a framework for translating Bengali to English. Context sensitive grammar rules are created for parsing. The top-down algorithm is used for parsing the sentences. We have proposed an algorithm for translating idioms in sentences. The proposed system is implemented and tested with about 15000 sentences. The performance analysis of the system gives 85.33% accuracy, which is quite satisfactory.**

Keywords—Bangla Machine Translator; Idioms; Bangla Language Processing (BLP); Left corner parsing algorithm.

I. Introduction

An Idiom is a commonly used word or sentence that implies something other than its metaphorical sense. Idioms convey a specific feeling and a specific tone for a language. Due to their common use, idioms can be recognized. Machine Translation (MT) relates to the application of computers, which is capable of translating the source language into target languages. This process generally does not have any human intervention. To translate vast quantities of data containing millions of words that might not be translated traditionally, MT techniques are used. This makes MT, a challenging task in the field of Natural Language Processing (NLP).

Native Bangla speakers are growing day after day by speaking and hearing idioms. It also implies for native English speakers. Idiom plays a vital role in the culture of different language speakers. In this modern age, it's very important to share knowledge and culture between different regions. But due to language barrier Bangladeshi's are not getting the advantage of learning the various culture. To overcome this barrier Bangla Language Processing can play an important role. Our Idiom translator will be able to help Bangali people to understand idioms in the English language, which will help them to adopt their culture and break the cultural barrier.

Generally, the MT model follows three main phases of parsing, transferring and generation. But our idiom translator follows four stages, which are idiom translator, parser, transfer, and generation. Idiom translator checks the idiom part in the sentence and translates it. Parser gathers the syntactic information of the sentence using Context Free Grammars (CFG). In the transfer stage, rules are transferred from source language to target language. And finally, the targeted sentence is generated in the generation stage. As idioms do not signify the literal meaning of the words used, it is hard to translate idioms from source language to target language.

The rest of the paper is organized as follows: Section II discusses related works. Methodology is discussed in Section III and it illustrates a sample following our proposed methodology. Section IV demonstrates the result and discussion and finally Section V refers the conclusion.

II. Releted Work

Research on the processing of natural language started in the 1950s. In the late 1980s, the first statistical machine translation systems were developed [1]. Till now many works are done in English language. Authors of [2] developed a Japanese-English machine translation system which was supported by the Japanese government's science and technology agency. The system applies many structural transformations during the transfer phase and generation phase to relieve the structural difference of the same contents and avoid ellipsis problems. Authors of [3] proposed an unsupervised Neural Machine Translation (NMT) system for translating English to German and German to English news.

Machine translation from Bangla language to other languages is in initial step now. Many works are done recently on Bangla to English or vice versa. A phrase-based Statistical Machine Translation (SMT) approach is proposed in [4]. In their work Out-of-Vocabulary (OOV) words are also handled. Authors of [5] proposed a rule-based transfer approach. They proposed an algorithm for searching the word from the lexicon and searching lexicon is made efficient by an intelligent integer based lexicon system. NLP techniques used to translate English to Bangla sentences in [6]. The context-free grammar used to validate the syntactical structure of a sentence and bottom-up approach is used to parse sentences. They used 50 sentences for every tense. In [7] they proposed a verb based machine translation approach for English to Bangla. They identified the main verb and make a simple form of English sentence. Then they easily translate it into Bangla. Authors of [8] also proposed context-sensitive grammar to translate Bangla to English. Bangla sentences including assertive, interrogative and imperative sentences. Set of context-sensitive grammar rules are proposed to translate imperative, optative and exclamatory Bangla sentences to English [9]. A new technique with a set of context-sensitive grammar rules is proposed to parse any Bangla sentences with imperative, optative and exclamatory Bangla sentences in [10] where moods got importance than the structure of sentence. They are generating a parse tree according to the sentences category.They used 400 sentences and got an accuracy of 81%. Authors of [11] work to find the appropriate verb according to the tense and subject. A procedure for finding semantically valid verb is proposed. They worked with verb root and different algorithms are proposed in this paper.

Maximum MT systems translate Bangla sentences to corresponding English sentences but we found only one of them includes idioms [12]. This paper presents, in addition to English, a multi lingual parallel idiom data set for seven Indian languages, and shows its relevance for two NLP applications. A set of CSG rules is proposed for our MT system to translate Bangla sentences with idioms to it's corresponding English sentence. Maximum work does not show the architecture of procedure of translation idioms and work with fewer data. In this system we proposed an architecture for translating sentences.

III. Proposed Methodology

In this propose system we have ten modules, the modules are idioms checker, idioms translator, tokenizer, rule generator, database, parser, target language rules, source language rules, machine translator and generator. Firstly, we consider a Bengali sentence "এই সমাজে বৃদ্ধ লোকেরা অচল পয়সা" as input of the system. Step by step procedure is given in Fig. 1.

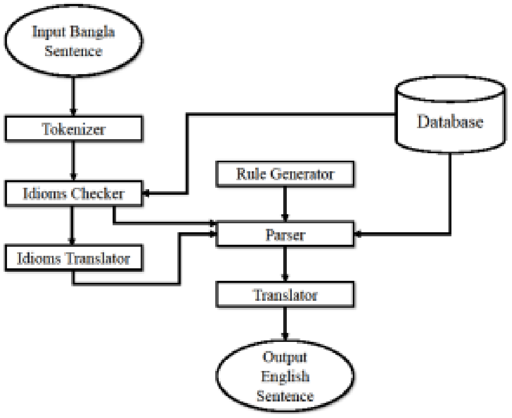


Fig. 1. Workflow of proposed system

A. Tokenizer

The main task of the tokenizer module is to split sentences into unit strings. It is like a database system of words with corresponding Parts of Speech (POS) tag. Suppose for the input sentence "এই সমাজে বৃদ্ধ লোকেরা অচল পয়সা ", the output will be like "এই","সমাজে", "বৃদ্ধ" ,"লোকেরা", "অচল", "পয়সা". After tokenizing the sentence, tokens will be going to idioms checker.

B. Idioms Checker

The main task of idioms checker is to check the idioms in the sentence by using Idioms checker algorithm which is Algorithm 1. In idioms dataset when w_i =অচল**, where** অচল also find in idioms dataset d_i, then it will find the next word w_{i+1} =পয়সা, then concat the string k = stringConcat(অচল, পয়সা). Now idioms d_i = is equal to k_{i+1} as idioms found in dataset so it will go to the next step Idioms translator if it does not find any term then it will concat the string up to i = 5 and then go to parser. If the sentence contains any idioms, then it will go to the idioms translator. For example,"এই সমাজে বৃদ্ধ লোকেরা অচল পয়সা" **as** "অচল পয়সা" is an idiom, it will go to the idioms translator module.

Algorithm 1: Algorithm for Idioms Checker

```
1. If $w_i$ is equal to split of $d_i$;
2. Find $w_{i+1}$;
3. Function mPairWord($w_1, w_2, .....w_n$);
4. k = function stringConcat($w_1, w_2, .....w_n$);
5. for $i = 0$ to idioms_dataset_length do
    if $k == d_i$ then
        go to 6;
        break;
    else
        go to 7;
    end
end
6. go to idioms Translator module;
7. go to Parser module;
```

C. Idioms Translator

This translator translates the idioms into its original meaning. As the idioms checker find that the sample input sentence has idioms"অচল পয়সা", after that this module translates the idioms into its corresponding meaning"মূল্যহীন". After translating idioms, it goes to parser module as shown in Fig 2.

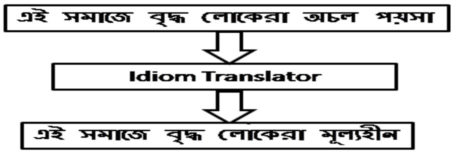

এই সমাজে বৃদ্ধ লোকেরা অচল পয়সা
Idiom Translator
এই সমাজে বৃদ্ধ লোকেরা মূল্যহীন

Fig. 2. Module of Idioms Translator

D. Database

Database module is just like a dictionary which contains the lexicon or token of a sentence and the related POS tag. For example, in this sentence the pos tag of corresponding words are,"এই" → **PN,** "সমাজে" → **N,** "বৃদ্ধ"→ **Adj,** "লোকেরা"→ **N,** "মূল্যহীন"→ **Adj**. In this system, it has another table which has a set of Bangla idioms and its meaning. Table I shows the Idioms_Table.

TABLE I
BANGLA IDIOMS TABLE

Idioms (d_i)	Meaning (m_i)
অচল পয়সা	মূল্যহীন
অকালকুষ্মাণ্ড	অপদার্থ
ইঁচড়ে পাকা	অকালপক্ব
উত্তম-মধ্যম	প্রহার
এলাহি কাণ্ড	বিরাট ব্যাপার

TABLE II
BANGLA CSGS RULES

Rule No	Bangla CSGs Rules
1	S → NP VP
2	NP → N (Biv) (Adj)
3	NP → N (Aux) (PP)
4	NP → NP NP
5	NP → (PN) N (Biv) (Adj)
6	NP → (Adj) N (Biv)
7	NP → (Qnt) (PP) N — PN
8	NP → N
9	PP → Null
10	V → Null
11	VP → V
12	VP → (Adj)
13	VP → (NP) VP
14	VP → V (Aux)
15	Adj → বৃদ্ধ, ভাল, অমূল্য, খারাপ, ...
16	PN → এই, আমি, আপনি, তুমি, ...
17	N → চোর, প্রহার, লোক, সমাজ,. . . . .
18	V → হয়, ছাড়ল, পড়া, খাওয়া, ...
19	Biv → টাকে, এরা, এ, ...
20	Aux → দিয়ে, পরে, করে, ...
21	Qnt → একটি, পাচটি, . . .

E. Rule Generator

The main purpose of the rule generator module is to generate the grammatical rules of Bangla sentences. For translating, the sentences, this module generates Context-Sensitive Grammar (CSG) rules. For this input sentence and this is built with the help of rules, these sentences need those rules NP → N (Biv) (Adj), S → NP VP, NP → (Qnt) (PP) N — PN for generating the parse tree. Sample CSG of Bangla simple sentences is listed in Table II.

F. Parser

Graphical view of the grammatical structure of the sentence is called the parse tree. Parser module helps to generate the parse tree of a sentence by using CSG rules and lexicon. We used left corner parsing algorithm to parse the sentence. This module generates the parse tree for the input sentence“এই সমাজে বৃদ্ধ লোকেরা মূল্যহীন” which is shown in Fig. 3.

G. Transfer

The task of the transfer module is to translate Bangla sentence to English language. The grammatical rule for transforming of grammar rule is listed in Table III. Using this grammar rules and transformation algorithm, we can get parse tree of English sentence, which is shown in Fig. 4.

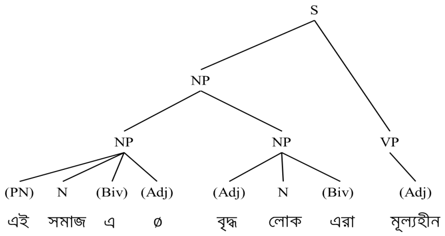


Fig. 3. Representation of Bangla parse tree

TABLE III
ENGLISH CSGS RULES

Rule No	English CSGs Rules
1	S → NP VP
2	S → VP NP
3	NP → NP NP
4	NP → Det N
5	NP → (PP) N (Adv)
6	NP → (PP) (PN) (Det) N
7	NP → Adj N
8	NP → Qnt N
9	NP → N
10	NP → PN
11	NP → (Aux) N
12	VP → V
13	VP → V (Adj)
14	VP → VP NP
15	VP → V (Gr) (N) (Adj)
16	VP → Aux V
17	N → thief, beating, society, person,
18	PN → this, that, I, She,...
19	V → release, are, like, eat, go,...
20	Adj → old, priceless, bad, good, ...
21	Aux → do, are, is,..
22	PP → in, on, to,..
23	Det → the, a, an,
24	Gr → ing

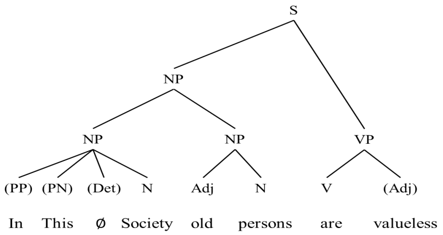


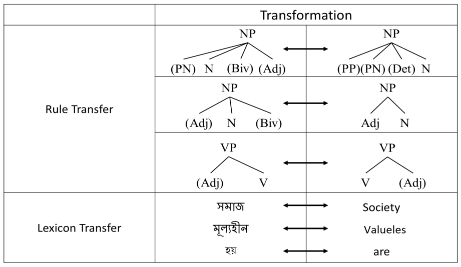
Fig. 4. Representation of English parse tree

The transformation process is divided into two part, rule transfer and lexicon transfer. The process of transforming grammar from source to target or from target to source language is shown in Table IV.

IV. EXPERIMENTAL RESULT

To assess the efficiency of our proposed system, we have evaluated the system with about 15000 distinct types of sentences with distinct sentence lengths. We collected these sentences from various books, websites, Bangla grammar books, Bangla text books etc.

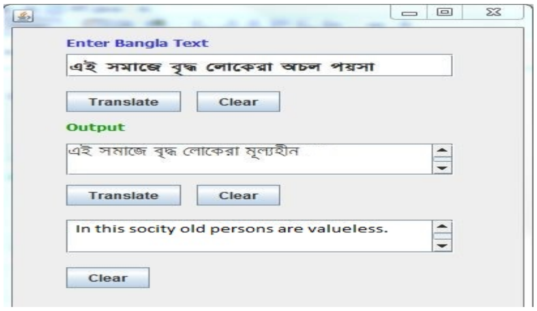
TABLE IV
TRANSFORMATION OF TARGET TO SOURCE OR VICE VERSA



	Transformation	
Rule Transfer	NP (PN) N (Biv) (Adj)	NP (PP)(PN) (Det) N
	NP (Adj) N (Biv)	NP Adj N
	VP (Adj) V	VP V (Adj)
Lexicon Transfer	সমাজ	Society
	মূল্যহীন	Valueles
	হয়	are

A. *Implementation*

For executing the system, we used, Windows 10 as the operating system, Java Swing to build the user interface, Java as the programming language, and NetBeans 8.2 as IDE. The snapshot of our implemented proposed MT system for the sentence "এই সমাজে বৃদ্ধ লোকেরা অচল পয়সা" with idioms is given in Fig. 5 where Google translator do not show the appropriate transformation, given in Fig. 6.

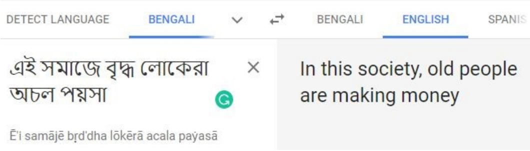


এই সমাজে বৃদ্ধ লোকেরা অচল পয়সা

এই সমাজে বৃদ্ধ লোকেরা মূল্যহীন

In this socity old persons are valueless.

Fig. 5. Translation of the sentence "এই সমাজে বৃদ্ধ লোকেরা অচল পয়সা"



এই সমাজে বৃদ্ধ লোকেরা অচল পয়সা

Ē'i samājē br̥d'dha lōkērā acala paẏasā

In this society, old people are making money

Fig. 6. Translation of the sentence "এই সমাজে বৃদ্ধ লোকেরা অচল পয়সা" in Google translator

B. *Accuracy Rate*

We observed that among 15000 sentences, a total of 12800 sentences were correctly translated with our proposed model. The accuracy rate is the ratio of the correctly translated sentences and the total number of sentences. Table V shows the accuracy rate for the sentences with different lengths. A graph of the system's accuracy rate vs. the sentence length is shown in Fig. 7. From this graph, we can observe that the accuracy rate is decreasing where the length of the sentences are increasing.

TABLE V
ACCURACY RATE OF DIFFERENT SENTENCES WITH DIFFERENT LENGTH

Sentences Length	No of input sentences	Correctly translated sentences	Overall accuracy (%)
3	3500	3300	94.24
4	3250	2850	87.69
5	3100	2650	85.48
6	2750	2150	78.18
7	2400	1850	77.08
Total	15000	12800	85.33

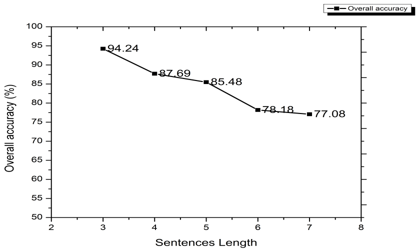


Fig. 7. Accuracy vs. word length graph

C. *Comparison Analysis*

Comparison with paper [12] of our proposed method is given in Table VI. Some parameters such as application, Emphasize, Feature, Accuracy etc. are shown in that comparison. In Table VI we can see that XML markup language used as feature in paper [12] where in our system rule-based approach is used which is more appropriate than XML markup language.

TABLE VI
COMPARISON BETWEEN PAPER [12] AND OUR PROPOSED SYSTEM

	Paper [12] [R. Agrawal, 2018]	Our proposed system
Application	MT, Sentimental Analysis	MT
Emphasize	Indian Languages	Only Bangla Language
Feature	XML markup	Rule Based
Accuracy	2.69% BLEU	score 85.33% for 0.015 million corpora
Dataset (Idioms)	2208 for 7 languages	986 for Bangla language

V. CONCLUSION

Aim of our paper is to translate different Bangla sentences containing idioms to its corresponding English sentences. The idea was to design a proper parsing technique to parse Bangla sentences with idioms. Our proposed algorithm is able to detect the idioms and translate it to its corresponding English meaning. The experimental result shows, our technique gives the accuracy of 85.33%. Our system might not get the exact parse tree for some sentences. To evaluate our implemented parsing model, we choose very simple and short Bangla sentences. It is possible to design a stronger parser for Bangla sentences to update CSG rules. These can be done by using semantic features for further research.

References

[1] Wikipedia. (2019) Natural language processing. [Online]. Available: https://en.wikipedia.org/wiki/Natural language processing

[2] M. Nagao, J. Tsujii, and J. Nakamura, "Machine translation from japanese into english," *Proceedings of the IEEE*, vol. 74, no. 7, pp. 993–1012, July 1986.

[3] M. Graça, Y. Kim, J. Schamper, J. Geng, and H. Ney, "The rwth aachen university english-german and german-english unsupervised neural machine translation systems for wmt 2018," in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 2018, pp. 377–385.

[4] M. Z. Islam, J. Tiedemann, and A. Eisele, "English to bangla phrase-based machine translation," in *Proceedings of the 14th Annual conference of the European Association for Machine Translation*, 2010.

[5] M. G. R. Alam, M. M. Islam, and N. Islam, "A new approach to develop an english to bangla machine translation system," *Daffodil International University Journal of Science and Technology*, vol. 6, no. 1, pp. 36–42, 2011.

[6] K. Muntarina, M. G. Moazzam, and M. A.-A. Bhuiyan, "Tense based english to bangla translation using mt system," *International Journal of Engineering Science Invention*, vol. 2, no. 10, pp. 30–38, 2013.

[7] M. Rabbani, K. M. R. Alam, and M. Islam, "A new verb based approach for english to bangla machine translation," in *2014 International Conference on Informatics, Electronics & Vision (ICIEV)*. IEEE, 2014, pp. 1–6.

[8] M. S. Arefin, L. Alam, S. Sharmin, and M. M. Hoque, "An empirical framework for parsing bangla assertive, interrogative and imperative sentences," in *2015 International Conference on Computer and Information Engineering (ICCIE)*. IEEE, 2015, pp. 122–125.

[9] T. Alamgir, M. S. Arefin, and M. M. Hoque, "An empirical machine translation framework for translating bangla imperative, optative and exclamatory sentences into english," in *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*. IEEE, 2016, pp. 932–937.

[10] T. Alamgir and M. S. Arefin, "An empirical framework for parsing bangla imperative, optative and exclamatory sentences," in *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, 2017, pp. 164–169.

[11] M. Haque and M. Hasan, "English to bengali machine translation: An analysis of semantically appropriate verbs," in *2018 International Conference on Innovations in Science, Engineering and Technology (ICISET)*. IEEE, 2018, pp. 217–221.

[12] R. Agrawal, V. C. Kumar, V. Muralidharan, and D. M. Sharma, "No more beating about the bush: A step towards idiom handling for indian language nlp," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.