Morphological Segmentation and Analysis of Bangla Text

G C Saha, Hasi Saha, Ruzinoor Che Mat, Nur Hossain Khan and Bappa Sarker

Abstract - This paper deals with lexicon and system development for word segmentation in Bangla language. Our goal in this paper is to develop a morphological segmentation algorithm that can work well for Bangla and to address the problem of unsupervised word segmentation across different languages. From a hand-corrected Bangla corpus, 5000 popular words were segmented into suffixes, prefixes and roots manually. These were the sample lexicon used as seed for next step. A system was developed using C language to automate the Segmentation process based on hand made lexical database. The System was evaluated on several pages of Bangla text and achieved a success rate of about 83.05%. In our observation the system will work with full success if twice the volume of lexicon database and this system may have a huge impact particularly to learn and use Bangla for the people which will enhance their socio-economic life greatly.

.

Index Terms- Bangla, Natural Language Processing, Lexicon, suffixes, prefixes and roots, morphological segmentation

1 INTRODUCTION

The expansion and research on computer Natural Language Understanding (NLU), has turned into a fascinating subject for some researchers throughout the most recent few decades. This has been further complimented by the advancement in speech recognition and Natural Language Processing (NLP) technologies; and the significant improvement of personal computer processing power and graphic technologies. The potential effect of common NLP has been broadly perceived since the earliest days of computers. Computer programmers for corpus linguistics and the need for further studies about how best to represent language varieties in a corpus. [**1**]. Thus one of the most widely used language Bangla, also known as Bengali, is the 4th most widely spoken language with more than 200 million speakers, most of whom live in Bangladesh and in the Indian state of West Bengal. Cutting edge Bangla morphology is exceptionally productive, particularly for verbs, with each root taking on 168 different forms. Bangla lexicon also has a very large number of simple and compound words, i.e., words that have more than one root, which can be made from any blend of nouns, pronouns and adjectives. While there are existing efforts at building a complete morphological analyzer for Bangla, all of these can only handle simple words with a single root.

From a research point of view, Bangla is highly inflectional, thus it can be required to posture similar difficulties to researchers in word segmentation just like Turkish and Finnish. Likewise the accessibility of a precise word segmentation algorithm for morphologically rich languages could considerably reduce the amount of annotated data needed to construct practical natural language. Hence, morphological analysis is a key component of NLP and computational linguistics. Along these lines, morphology is the branch of linguistics that studies patterns of word formation within and across languages, and attempts to plan rules that model the learning of the speakers of those natural languages. Hence, NLP is the computerized approach to analyzing text that depends on both an arrangement of theories and an arrangement of technologies. And, being an exceptionally area of research and innovative work, there is not a solitary settled upon definition that would satisfy everyone, but there are some aspects, which would be part of any knowledgeable person's definition. The historical backdrop of morphological examination goes back to the antiquated Indian language specialist Pāṇini, who formulated the 3,959 rules of Sanskrit morphology in the text *Aṣṭādhyāyī* by using a Constituency Grammar. The Greco-Roman grammatical tradition also engaged in morphological analysis. According to Badruddoza [2] an online Bangla handwritten recognition system was reported that uses neural network for feature selection and extraction, and achieves a recognition rate about 90%.

Morphology is the distinguishing proof, analysis and depiction of the structure of words (words as units in the lexicon are the subject matter of lexicology) or the division of a word into smaller sub-parts, or morphemes. Morphological analysis and segmentation is the task of

- *G C Saha is with the Department of Computer Science & Information Technology,Bangabandhu Sheikh Mujibur Rahman Agricultural University,Gazipur,Bangladesh, 1706. E-mail: gcsaha@bsmrau.edu.bd*
- *Hasi Saha is with the Department of Computer Science & Information Technology,Hajee Mohammad Danesh Science & Technology University,Dinajpur,Bangladesh.E-mail: hasi.cse3@gmail.com*
- *Ruzinoor Che Mat is with the School of Multimedia Technology & Communication, Universiti Utara Malaysia, Malaysia, Kedah 06010. E-mail: ruzinoor@uum.edu.my*
- *Nur Hossain Khan is with the Department of Computer Science & Engineering, Islamic University,Kushtia, Bangladesh. E-mail: nur_cse_iu@yahoo.com*
- *Bappa Sarker is with the Department of Computer Science & Engineering, Islamic University,Kushtia,Bangladesh. E-mail: bappacse07@yahoo.com*

segmenting a word into morphemes (i.e. prefixes, suffixes and roots), the smallest meaning-bearing elements of natural languages. For example, the English word "unforgettable" is divided into 3 morphemes, i.e. "un", "forget", and "able". Similarly, the Bangla word অনাধুনিকতার ("anAdUnIktAr") is divided into "an" (PREFIX), "AdUnIk" (ROOT), "tA" (SUFFIX) and r" (INFLECTION). Thus morphology is the identification, analysis and description of the structure of words (words as units in the lexicon are the subject matter of lexicology). While words are by and large accepted just like the littlest units of Natual Language, it is clear that in most (if not all) languages, words can be identified with different words by guidelines. For example, English speakers recognize that the words *dog*, *dogs*, and *dog catcher* are closely related.Therefore, it is essential to take into account a morphological analysis of Bangla Words for the Universal Network Language (UNL) system to include Bangla as a member of UNL. At the some previous decades there has been a considerable amount of work on knowledge-based morphological analysis for none of these knowledge-based analyzers have been empirically evaluated [3, 4, 5, 6]. In order to enable a virtual character to interact with humans via language, the character should have the capability of understanding humans through speech recognition, natural language understanding, and communication via natural language generation and speech [7]. In a subsequent paper, Goldsmith [8] adopts the Minimum Description Length (MDL) approach and provides a new information-theoretic compression system that gives a measure of the length of the morphological grammar. He applies his algorithm to English and French and reports accuracies of 82.9% and 83.3% respectively. He also groups together the possible suffixes for each stem, and introduces the signature paradigm that is helpful for determining syntactic word classes (*i.e.*, part-of-speech classes). Motivated by Goldsmith, Creutz [**8**] and Creutz and Lagus [9] propose a probabilistic maximum *a* posteriori formulation that uses prior distributions of morpheme length and frequency to measure the goodness of an induced morpheme. They work on English and Finnish (a highly agglutinative language) and report better accuracy than Goldsmith's Linguistica morphological parser. The last approach, introduced by by Freitag [10], first automatically clusters the words using local co-occurrence information and then induces the suffixes according to the orthographic dissimilarity between the words in different cluster

In spite of the fact that exceptionally fruitful, knowledge based ways to deal with word division work by depending on manually outlined heuristics, which require a lot of linguistic expertise and are also time-consuming to construct. As a result, research in morphological analysis has exhibited a shift from knowledge based approaches to unsupervised approaches. Unsupervised word segmentation is ordinarily made out of two stages: (1) a morpheme induction step in which morphemes are automatically induced from a hand-handled database consisting of words taken from a large, un annotated corpus, and (2) a segmentation step in which a given word is segmented by morphological segmental algorithm which based on induced morphemes. Unsupervised word segmentation has achieved considerable success [11, 12 and 13]. For instance, Schone and Jurafsky report F-scores of 88%, 92%, and 86% on English, German, and Dutch word segmentation, respectively. It will also push the Bangla language identity on the global nation. This paper aims to develop a segmentation system of Bangla text plays an important role in morphological recognition because it allows the recognition system to classify the characters more accurately and quickly.

2 Techniques Adopted in the Implemented System

For instance, builds up a system for distinguishing morpheme that checks whether the number of different letters following a sequence of letters exceeds some given threshold that rely on upon successor and antecedent frequencies to identify morpheme database. We then manually segmented each of the 5000 hand-segmented Bengala words as Prefix+Root or Root+Suffix to develop the database. We use here corpus of approximately 5000 words, which is very small compared to the number of word types typically seen in existing literature on unsupervised morphological induction. This segmentation is a by first inducing a list of most frequent morphemes and then using those morphemes for word segmentation. The goal is to find a set of morphemes such that when each word in a given corpus is segmented according to these morphemes, the total length of an encoding of the corpus is minimized.

3 Database Development

3.1 Simple affixes, roots and suffixes generation

At first we take a cleaned Bangla corpus from which various individual Bangla words are taken randomly. Then those words are segmented by hand-held into roots, affixes and suffixes which produce relative prefixes lexicon, suffixes lexicon and roots lexicon respectively is shown below in Table 1.

Table 1**:** Simple root, prefixes and suffixes generation

Main Word	Root	Prefixes	Suffixes
থইথই	**থই**		
অথই	**থই**	**অ**	
থানার	থানা		**ও**

3.2 The Basic Morpheme Algorithm

We use here machine independent high level programing language C for developing the segmentation system and further for demonstrate the segmentation procedure for 5000 hand corrected Bangla words. Our unsupervised segmentation algorithm is composed of two steps: (1) inducing *prefixes*, *suffixes* and *roots* from a corpus that consists of words taken from a large corpus, and (2) segmenting a word using these induced morphemes. This section describes our *basic* morpheme induction method.

4 WORD SEGMENTATION

Pseudocode algorithm for Simple word segmentation:

Start the program
Open source file in read mode (fp1)
Open file (fp2) in read mode
Open output file (fp) in write mode
/* store affixes, suffixes & roots */
Read one character from file (fp1), store in variable str2
Check str2! = ###
Read one character from file (fp2), store in variable str1.
Check str1! = ### Then
Loop until p! =0 & q!=0
(a) Compare str11 with str2, store value in variable in Ptr.
(b) Check Ptr!=Null
(c) Calculate length of str1 and str2
(d) Segment str1
Store segment value by using file (fp)
Close (fp)
close (fp1)
close (fp2)
End of program.

We have executed the above morphological analyzer for both simple and compound words which is based on two-level morphology. We have used both simple and compound-words found from the popular daily Bangla newspapers to produce our test cases and got expected correct result. A flowchart showing controls the flow of execution based on a condition can be referred at figure 1. Therefore it will work for any given inflectional compound word whether it is in our test cases or not.

This clearly is a somewhat coarse analysis of morphotactic structure, and as such greatly over recognizes. For example it recognizes both hAtCIlAm (হাটছিলাম) and hAtc~CIlAm (হাটছিলাম).
For example if a word is given like this:
hEtECIlAm = hAt + EC + Il + Am
For example
অনাধুনিকীকরণের
= অন+আধুনিক+করণ+এর
= Prefix+ Nroot + Suffix + Suffix
Let us say that we want to recognize the word "krCil".

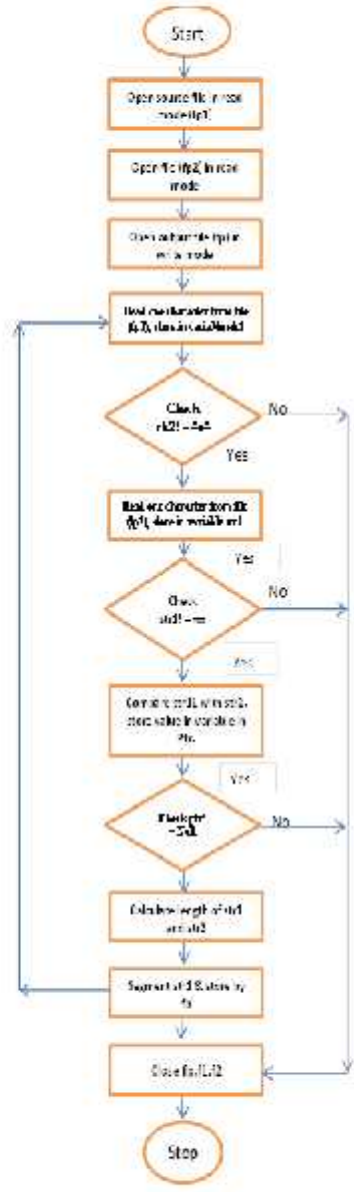

Figure 1: Flowchart of simple word segmentation algorithm

Presently the corresponding information for the word in the verb root lexicon is as per the following:
\lf kr
\lx VROOT
\alt Aspect
\gl kr
The above information indicates that the word kr is a root word, the following spot that we need to go is the Aspect and the lexical entry for the word is kr. From here we reach the verb suffix file. The place we hit is the Aspect entry. This section determines the strained of a word in connection to its suffix.

5 EVALUATION

Now, let us evaluate our segmentation algorithm.

5.1 Experimental Setup

Creating corpus. The corpus from which we extract our database contains one month of news taken from the Bangla newspaper "Ittefaq". We then pre-process each of these articles by tokenizing it and removing punctuations and other unwanted character sequences (such as .*** ###.). The remaining words are then used to create our database, which consists of 5000 distinct words. Not at all like morphological analysis for many other languages, however, we do not take the conventional step of removing extra parts from our database, because we do not have a entity identifier for Bangla.

Test set preparation. To build our test set, we randomly choose 5000 words from our database that are at least 5-characters long. We impose this length restriction when selecting our test cases simply because words of length one or two do not have any morphological segmentation in Bangla. We then manually remove the proper roots, affixes and suffixes with mistakes from the test set before giving it to two of our linguists for hand-segmentation. In the absence of a complete learning based morphological parsing tools and a hand-tagged morphological database for Bangla, our linguists had to depend on the Bangla database for clarifying our test cases. One example of such word is বিরুদ্ধ (bIrUd~D), whose actual segmentation is বি+র +রু (□) (bI+rUd+k~T (T)) which is tough to obtain. However, if the meaning of a segmented word differs from that of the original word, then we simply treat the original word as a root (i.e. the word should not be segmented at all). Words that fall inside this class incorporate প্রধা , □ □□□□ and প্রতি□□□□. After all the words have been manually segmented, we remove those for which the two linguists produce inconsistent segmentations. The subsequent test set contains a few words.

5.2 Experimental Results

To evaluate morphological system performance, a pre processed test data set (about 1000 correct spelled words) from hand made lexical database was run through the developed analyzer and the result was compared to correct recognition words which produced correct number of words 896 and the recognition rate was 89.9% as well. The evaluation was again conducted across several wrong spelled words (about 1000 wrong spelled words) and in the same way number of correct words with recognition rate was 765 and 76.5% respectively .
Table 2 shows the accuracy and performance result on the prepared test sets, and shows it's total morphological system performance as 83.05% .

Table 2: Morphological System Performance

Test words	Number of test words (N)	Correct Recognition (n)	Recognition Rate(%) $= n * 100/N$
Correct spelled Words	1000	896	89.6
Wrong spelled Words	1000	765	76.5
Total Morphological System Performance			83.05%

5.3 Discussion and Error Analysis

As a component of the analysis of our word segmentation algorithm, we are interested in testing whether it can correctly segment complicated test cases. Encouragingly, our system successfully segments complex Bangla words like. □□□□□□□ (dUlIyECIl) as .dUl+IyE+CI+l., as well as multi-root words like বিনোদনকেন্দ্রগুলো (bInOdnkEndRgUlOo), whose correct segmentation is .bInOd+n+kEndR+gUlO+o.. Even more interestingly, it correctly parses English words, which are widely used in the sports section of the newspaper. For example, words like বোলিং (blIng) and ফাইনালিস্ট (FAinAlIS~t) are correctly segmented into bl+Ing. And FAinAl+IS~t. It randomly specifying that the compounding nature of Bangla and the influence of foreign languages have introduced into our repository a lot of new words, whose presence increases the difficulty of the segmentation task. By the way, our word segmentation system manages to stem those words correctly. Thus our developed morphological segmentation systems achieve a decent performance of 83.05%.

6 CONCLUSION AND FUTURE WORK

We have presented a new morphological analyzer for Bangla word segmentation that, when evaluated on a set of 5000 human-corrected Bangla words, substantially outperforms database. Analysis reveals that our novel uses of segmentation algorithm along with our proposed technique for the detection of prefix, root and suffix, have contributed to the superior performance of our algorithm. In future work, we plan to investigate whether our algorithm can be improved by incorporating automatic irregular word form detection and using automatically acquired information about the semantic relatedness between word pairs. The System was assessed on a few pages of Bangla text content

from hand corrected lexicon and made a progress rate of around 83.05%. The developed sample Bangla lexicon and a good morphological segmentation system which will be helpful for spelling and grammar checking, speech reconstruction, speech generation, topic detection, message understanding and many other related topics which will immensely help the students, researchers and other people in our society. In adition, we plan to construct a Part-Of-Speech (POS) tagger for Bangla that adventures the morphological information gave by our framework. This contrasts with existing work on POS tagging for Bangla languages, where POS taggers are commonly built by using information provided by the morphological word segmentation system. Hopefully our effort here will help implementing a complete-morphological analyzer for Bangla in future.

ACKNOWLEDGMENT

The authors wish to thank Sir Dr. *Md. Farukuzzaman Khan*, Professor, Department of Computer Science and Engineering, Islamic University & Former Dean, Faculty of Applied Science and Technology, Islamic University, Kushtia for his patronage to dissemination of knowledge, his valuable instructions, able guidance and keen interest throughout the research. The author is also grateful to Dr.Ruzinoor Che Mat, School of Multimedia Technology & Communication, Universiti Utara Malaysia, Malaysia, Kedah 06010 for his encouragement, precious comments and valuable suggestion in preparation of this work.

REFERENCES

[1] Conrad, Susan. " Corpus linguistic approaches for discourse analysis." *Annual Review of Applied Linguistics,2002* Mar 22(1):75-95.

[2] Badruddoza, M. "Recognition of Bangla hand written letters using self-organizing map (SOM)". Proceedings of 6th International Conference on Computer and Informati on Technology (ICCIT), 357-360, 2003.

[3] Samit Bhattacharya, Monojit Choudhury, Sudeshna Sarkar and Anupam Basu. "Inflectional morphology synthesis for Bengali noun, pronoun and verb systems". In Proceedings of the National Conference on Computer Processing of Bangla (NCCPB05), pp. 34 – 43, 2005

[4] Sajib Dasgupta and Mumit Khan. "Feature Unification for Morphological Parsing in Bangla." In the Proceedings of 7th ICCIT, Bangladesh, 2000

[5] Dash NS. The Morphodynamics of Bengali Compounds decomposing them for lexical processing. Language in India (www. languageinindia. com), 6(7), 2006.

[6] Dey K, Bhattacharyya P. Universal Networking Language based analysis and generation of Bengali case structure constructs. Res. Comput. Sci., 12, pp. 215-29, 2005.

[7] Schone P, Jurafsky D. Knowledge-free induction of inflectional morphologies. In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies, pp. 1-9, 2001.

[8] Goldsmith J. Unsupervised learning of the morphology of a natural language.Computational linguistics. UniversityofChicago.1997

[9] Creutz M, Lagus K. "Inducing the morphological lexicon of a natural language from unannotated text" In Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Vol. 1, No. 106-113, pp. 51-59, 2005.

[10] Creutz M, Lagus K. "Induction of a simple morphology for highly-inflecting languages." In Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology: Current Themes in Computational Phonology and Morphology, Association for Computational Linguistics,pp. 43-51, 2004.

[11] Creutz M, Lagus K. "Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0." Helsinki University of Technology; 2005 Mar.[11] Dasgupta S, Ng V. "Unsupervised word segmentation for Bangla". Proceedings of ICON, pp. 15-24, 2007.

[12] John Goldsmith. "Unsupervised learning of the morphology of a natural language." In Computational Linguistics ,Vol. 27 no. 2, pp. 153-198, 2001

.

[13] Patrick Schone and Daniel Jurafsky. " Knowledge-free induction of inflectional morphologies". In Proceedings of the Second Meeting of the North American Chapter of Association of Computational Linguistics (NAACL), ,pp. 183-191, 2001.

G C Saha is an Assistant Professor at the Department of Computer Science & Information Technology at Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur-1706, Bangladesh.He has been as the faculty member of BSMRAU since December 2011. His research interest is in the area of 3D GIS, remote sensing application, machine learning and visualization. He is a Computer Science & Engineering graduate from Islamic University; Kushtia, has worked on many ICT based projects that involve innovation and development in the field of Information Technology.

Hasi Saha is an Assistant Professor at the Department of Computer Science & Information Technology, Hajee Mohammad Danesh Science & Technology University, Dinajpur, Bangladesh. She received her BSc in Computer Science &

Engineering at the same University and MSc in IT by sresearch in 2015 from Dhaka University. Dhaka, Bangladesh. She has involved in Computer Science field since 2007 and her research interests include password based authentication and system learning.

Ruzinoor Che Mat is a Senior Lecturer at the School of Multimedia Technology and Communication, Universiti Utara Malaysia,UUM. His research areas include reverse engineering, 3D GIS, terrain visualization, remote sensing application, virtual reality, computer graphics and visualization. He received BEng (Hons.) Electrical and Electronic Engineering from Coventry University, UK, MSc. Computer Graphics and Virtual Environment from University of Hull, UK and PhD in GIS and Geomatic Engineering from Universiti Putra Malaysia.

Nur Hossain Khan obtained his bachelor and Masters' degree in Computer Science & Engineering from Islamic University, Kushtia. Bangladesh. Currently he is serving as Assistant Maintenance Engineer (Assistant Director) at Bangladesh Bank. His research interest includes Natural language processing and machine learning.

Bappa Sarker is a Lecturer at the Department of Computer Science & Engineering under Islamic University, Kushtia, Bangladesh. He also completed his BSc and MSc in Computer Science & Engineering (CSE) at the same University. His research interest is in the area of Natural Language Processing and Machine Learning. He has experience in the field of Computer Science.

.