Morphological Segmentation and Analysis of Bangla Text

Gonesh Chandra Saha
Bangabandhu Sheikh Mujibur Rahman Agricultural University

Hasi Saha
Hajee Mohammad Danesh Science and Technology University

Bappa Sarkar
Islamic University (Bangladesh)

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 objective in this article is to build up a morphological segmentation algorithm that can function admirably for Bangla and 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 a seed for next step. A system was developed using C language to automate the Segmentation process based on hand made a 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 colossal effect especially to learn and utilize Bangla for the general population which will upgrade their socioeconomic life significantly.

.

Index Terms— Bangla, Natural Language Processing, Lexicon, suffixes, prefixes, and roots, Morphological segmentation

1 INTRODUCTION

The expansion furthermore, explore on PC Natural Language Understanding (NLU), has turned into a fascinating subject for some researchers throughout the most recent barely any decades. This has been additionally complimented by the headway in speech acknowledgment and Natural Language Processing (NLP) advances, 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 languages Bangla, otherwise called Bengali, is the fourth most generally talked with more than two hundred million speaking language, the greater part of whom live in the Indian territory of West Bengal and in Bangladesh. Bleeding edge Bangla morphology is especially profitable with each root going up against 168 unique structures, especially for verbs. Bangla dictionary additionally has countless compound words, for instance, words with multi roots, that can be produced using any mix of adjectives, pronouns, and nouns. While any mix of adjectives, nouns, and pronouns. In the meanwhile the existing endeavors at construction an entire analyzer of morphology for Bangla, that can just deal with basic words with a solitary root.

From a research point of view, Bangla is highly inflectional, thus it can be required to posture similar difficulties to scientists in word segmentation simply like Finnish and Turkish. Likewise, the accessibility of a precise word division algorithm for morphologically rich dialects could significantly diminish the measure of commented on information expected to build handy natural language. Subsequently, the morphological investigation is a key part of NLP and computational etymology. Thusly, morphology is the branch of phonetics that reviews examples of word arrangement inside and crosswise over language, and endeavors to design decides that model the learning of the speakers of that natural language. 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 recorded backdrop of the examination of morphological backpedals to the out of date Indian dialect expert Pāṇini, who nitty gritty the 3,959 standards of Sanskrit morphology in the substance Aṣṭādhyāyī by using a constituency grammar. The Greco-Roman syntactic convention additionally occupied with morphological analysis. As indicated by Badruddoza [2] an online Bangla written by hand acknowledgment system was accounted for that utilizations neural system for feature determination and extraction and accomplishes an acknowledgment rate around 90%.

- *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. email: 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*

Morphology is the distinguishing proof, examination, and analysis of the structure of (words as units in the lexicon are the topic of lexicology) or the word segmentation into morphemes or little subparts. Morphological division and the examination is the errand of separating a word into morphemes (i.e. roots, prefixes, and suffixes), the littlest importance bearing components of common natural language. For instance, The “unforgettable” English word is isolated into three morphemes, i.e. "un", "forget", and "able". Thus, the word অনাধুনিকতার ("anAdUnIktAr") in Bangla is isolated into "a" (PREFIX), "AdUnIk" (ROOT), "tA" (SUFFIX) and "r" (INFLECTION). Along these lines, morphology is the analysis, identification, and depiction of the structure of (words as units in the lexicon are the topic of lexicology). While words are by and large accepted just like the littlest units of Natual Language, obviously in most (if not all) dialects, words can be identified with different words by guidelines. For example, English speakers see that the words dog, dogs, and dog catcher are about related. Therefore, it is basic to consider a morphological analysis for Bangla Words for the UNL system to incorporate Bangla as an individual member of UNL. At 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]. Keeping in mind the end goal to empower a virtual character to associate with people by means of language, the character ought to have the ability to understand people through discourse recognition, understanding of natural language comprehension, and correspondence by means of speech and natural language era[7]. In a consequent paper, Goldsmith [8] embraces the Minimum Description Length (MDL) approach and gives another data theoretic pressure framework that stretches a quantity of the morphological length of the linguistic use. The researcher applies that calculation to French and English and reports correctnesses of 83.3% and 82.9% separately. He likewise bunches together the conceivable stem suffixes and presents the worldview signature that is useful for deciding syntactic word modules (i.e., grammatical feature groups). Spurred by Creutz, Goldsmith [8] and Lagus & Creutz [9] suggested a probabilistic most extreme definition that utilizations earlier dispersions length of morpheme and recurrence to gauge the decency of an actuated morpheme. They deal with Finnish and English (an exceedingly dialect) what's more, report preferable precision over Linguistica morphological parser of Goldsmith. The successive approach, presented by Freitag [10], first naturally bunches the words utilizing nearby co-event data and after that incites the suffixes as indicated by the uniqueness of orthographic among the words in various 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 great deal of linguistic based expertise and are likewise tedious to develop. Accordingly, a study in the morphological analysis has shown a move from knowledge based approaches to unsupervised approaches. Unsupervised word segmentation is ordinarily made out of two stages: (1) a morpheme enlistment venture in which morphemes are consequently induced from a hand-handled care of database comprising of words are taken from an extensive, un clarified mass corpus, and (2) the division venture where in a specified word is sectioned by morphological segmental algorithm which based on induced morphemes. Unsupervised word division has made extensive progress [11, 12 and 13]. For example, Jurafsky and Schone account F-scores of 92%, 88%, and 86% on German ,English, and Dutch word division, separately. It then may likewise impulse the Bangeli language character through the worldwide country. This study findings expects to build up a segmentation system of Bangla content assumes a vital part in morphological recognition since it enables the acknowledgement system to characterize the typescripts rapidly and all the more precisely.

2.TECHNIQUES ADOPTED IN THE IMPLEMENTED SYSTEM

For instance, builds up a system for distinguishing morpheme that exmines whether a majority of various characters following a succession the characters crosses approximately assumed edge that relies on upon successor and antecedent frequencies to identify morpheme database. We then manually segmented each of the 5000 hand-segmented Bengala words as Root + Prefix or Suffix + Root to develop the database. We use here corpus of approximately 5000 words, which is little contrasted with a lot of word sorts commonly observed in existing writing on unsupervised morphological acceptance. This segmentation is a by first actuating a rundown of most continuous morphemes and afterward utilizing those morphemes for word division. The objective is to locate a set of morphemes with the end goal that once individually word in a agreed corpus is portioned by these morphemes, the aggregate extent of an encrypting corpus is reduced.

3. DATABASE DEVELOPMENT

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

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 hands corrected Bangla words. Our unsupervised division algorithm is made out of 2 stages: (1) actuating roots, suffixes and prefixes from a corpus that comprises of words taken from a big corpus, and (2) dividing a word utilizing these instigated morphemes. This area portrays our technique for the acceptance of fundamental morpheme.

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 then have executed the above tested morphological segmentation analyzer for both compound and basic words which depends on two-level morphology. We have utilized both basic and compound-words found from the prominent day by day Bangla daily newspapers to deliver our experiments and got expected the right outcome. A flowchart showing controls the flow of execution based on a condition is referred to below figure 1. In this way, it will effort for any prearranged inflectional multiple word whether it is in our experiments or not.

This clearly is a somewhat coarse examination of the morphotactic structure, and thusly enormously finished perceives. For instance, it perceives both hAtCIlAm（হাটছিলাম）and hAtc~CIlAm（হাটছিলাম）.

For instance if a word is given this way:

hEtECIlAm = hAt + EC + Il + AmFor example
অনাধুনিকীকরণের

= অন+আধুনিক+করণ+এর

Figure 1: flowchart of simple word segmentation algorithm

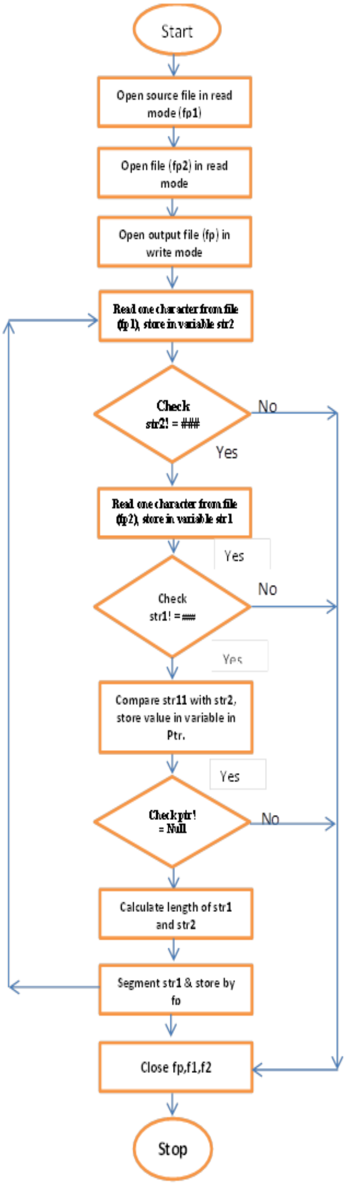
= Prefix+ Nroot + Suffix + Suffix

5. EVALUATION

Now, let us evaluate our segmentation algorithm.

5.1 Setup Experiment

Creating corpus. The corpus beginning with which we remove our database comprises one month of Bangla news reserved from the Bangla daily paper "Ittefaq". We then pre-process each of these trainings corpus by tokenizing it and evacuating punctuations and other undesirable character successions, (for example,*** ###.). The rest of the words are at that period utilized to make our database, which comprises of 5000 unmistakable words. Not in any way like a morphological examination for some different language, be


Start
Open source file in read mode (fp1)
Open file (fp2) in read mode
Open output file (fp) in write mode
Read one character from file (fp1), store in variable str2
Check str2! = ###
No
Yes
Read one character from file (fp2), store in variable str1
Yes
Check str1! = ###
No
Yes
Compare str11 with str2, store value in variable in Ptr.
Yes
Check ptr! = Null
No
Calculate length of str1 and str2
Segment str1 & store by fo
Close fp,f1,f2
Stop

that as it may, we don't make the ordinary stride of expelling additional parts from our database, since we don't have a substance identifier for Bangla.

Preparing test set. To fabricate our test set, we haphazardly pick 5000 words from our database that are no less than 5-characters in length. We force this length confinement while choosing our experiments basically in light of the fact that expressions of length maybe a couple don't have any morphological division in Bangla. We at that point physically remove the best possible prefix, suffixes, and roots with errors from the test set before offering it to two of our language specialists for hand-division. Without an entire learning based morphological parsing instruments and a hand-labeled morphological database for Bangla, our language specialists needed to rely upon the Bangla database for clarifying our experiments. One case of such word is বিরুদ্ধ (bIrUd~D), whose real division is বি+রুধ+ক্ত(ত) (bI+rUd+k~T (T)) that is difficult to get. Be that as it may, if the importance of a sectioned word varies from that of the first word, at that point we, regard the first root word (i.e. the word ought not toward be portioned by any stretch of the imagination). Words that fall inside this class consolidate প্রধান,আবেদন and প্রতিবেদন. After every one of the words has been physically divided, we remove those for which the two language specialists create conflicting segmentation. The ensuing test set contains a couple of 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 a correct number of words 896 and the recognition rate was 89.9% as well. The evaluation was again conducted across several wrongs spelled words (about 1000 wrong spelled words) and in the same way number of correct words with the recognition rate was 765 and 76.5% respectively.

Table 2 shows the accurateness and performance result on the prepared test sets and shows it's total morphological system performance as 83.05%.

Table2**:** 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 examination of our word division algorithm, we are occupied with testing whether it can accurately fragment convoluted experiments. Reassuringly, our system effectively sections complex Bangla words like. the system effectively fragments complex Bangla words like. দুলিয়েছিল (dUlIyECIl) as .dUl+IyE+CI+l., and also multi-root words like বিনোদনকেন্দ্রগুলো (bInOdnkEndRgUlOo), whose correct segmentation is. bInOd+n+kEndR+gUlO+o. Considerably more strangely, it effectively parses English words, which are broadly utilized as a part of the sports segment of the daily newspaper. For instance, words like বোলিং (blIng) and ফাইনালিস্ট (FAinAlIS~t) are effectively fragmented into bl+Ing. Also, FAinAl+IS~t. It haphazardly determining that the intensifying idea of Bangla and the impact of outside language have brought into our repository a considerable measure of new words, whose nearness builds the trouble of the segmentation errand. Coincidentally, our word division system figures out how to stem those words effectively. Thus our developed morphological segmentation systems achieve a decent performance of 83.05%.

6. CONCLUSIONS AND FUTURE WORK

We have introduced another morphological analyzer for Bangla word segmentation that, at the point when assessed on an arrangement of 5000 human-corrected Bangla words, substantially outperforms database. The analysis uncovers that our novel employments of segmentation algorithm along with our proposed technique for the detection of prefix, root, and suffix, have added to the predominant execution of our calculation. In future work, we plan to investigate whether our algorithm can be improved by incorporating automatic irregular word form detection and using naturally procured data about the semantic relatedness between word sets. 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 accommodating for spelling and language structure checking, speech reproduction, speech era, point discovery, message understanding and numerous other related themes which will tremendously help the researchers, students and other individuals in our society. In addition, our strategy to construct a Part-Of-Speech (POS) tagger for Bangla that adventures the morphological data gave by our structure. This appears differently in relation to existing work on POS labeling for Bangla language, where POS taggers are generally worked by utilizing data gave the morphological word division framework. Ideally, our exertion here will help to actualize a total 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 the 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 for this work. This research was partially upheld by ICT Division, Ministry of Posts, Telecommunications and IT, the government of the People's Republic of Bangladesh.

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 Information 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. language in India. 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 Professor 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 for 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 research 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 the University of Hull, UK and Ph.D. 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.
.