STATISTICAL SANDHI SPLITTER FOR BENGALI COMPOUND WORDS

Thesis

Submitted In Partial Fulfillment Of The Requirement For The Degree of

MASTER OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

BY

SOUMYABRATA CHATTERJEE

University Roll Number: 002010502029

Examination Roll Number: M4CSE22029

Registration Number: 154153 of 2020-2021

Under The Guidance Of

PROF. DIGANTA SAHA

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

FACULTY OF ENGINEERING & TECHNOLOGY

JADAVPUR UNIVERSITY, KOLKATA

132, Raja Subodh Chandra Mallick Road

Jadavpur, Kolkata, West Bengal, 700032

JUNE, 2022

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

CERTIFICATE OF RECOMMENDATION

This is to certify that the dissertation titled STATISTICAL SANDHI SPLITTER FOR BENGALI COMPOUND WORDS was completed by SOUMYABRATA CHATTERJEE, University Roll No: 002010502029, Examination Roll Number: M4CSE22029, University Registration No: 154153 of 2020-21, under the guidance and supervision of Prof. Diganta Saha, Department of Computer Science and Technology, Jadavpur University. The findings of the research detailed in the thesis have not been incorporated into any other work submitted for the purpose of earning a degree at any other academic institution.

_____

Prof. Diganta Saha
Department of Computer Science & Engineering
Jadavpur University

COUNTERSIGNED BY

_____

Prof.(Dr.)Anupam Sinha
Head of The Department
Department of Computer Science
and Engineering
Jadavpur University

COUNTERSIGNED BY

_____

Prof. Chandan Mazumdar
Dean, FET
Department of Computer Science
and Engineering
Jadavpur University

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

CERTIFICATE OF APPROVAL

This is to certify that the thesis entitled STATISTICAL SANDHI SPLITTER FOR BENGALI COMPOUND WORDS is a bonafide record of work carried out by SOUMYABRATA CHATTERJEE in partial fulfilment of the requirements for the award of the degree Master of Engineering in Department of Computer Science and Engineering, Jadavpur University during the period of June 2021 to May 2022 (3^{rd} & 4^{th} Semester). It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn there in but approve the thesis only for the purpose for which it has been submitted.

_____

Signature of Examiner

Date:

_____

Signature of Supervisor

Date:

DECLARATION

I certify that,

a) The work STATISTICAL SANDHI SPLITTER FOR BENGALI COMPOUND WORDS contained in this report has been done by me under the guidance of my supervisor.
b) The work has not been submitted to any other Institute for any degree or diploma.
c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I havetaken permission from the copyright owners of the sources, whenever necessary.

Soumyabrata Chatterjee
Master of Engineering
Roll No.: 002010502029
Registration No.: 154153 of 2020-21
Department of Computer Science and Engineering
Jadavpur University, Kolkata

ACKNOWLEDGEMENT

First and foremost, I want to express my gratitude to God Almighty for providing me with the strength, wisdom, and capability to go on this amazing adventure and to continue and successfully finish the embodied research work. I'd like to thank Professor Diganta Saha of Department of Computer Science and Engineering at Jadavpur University for his excellent assistance, consistent support, and inspiration during the course of my dissertation. I owe Jadavpur University a great debt of gratitude for providing me with the chance and facilities to complete our thesis.

I am grateful to every one of the teaching and non-teaching personnel whose assistance has made the trip during my research time much easier. I would like to thank my seniors, and my friends for providing me with regular encouragement and mental support throughout our effort.

Last but not the least, my family deserves great recognition. There are no words to express my gratitude to my mother and father for all of the sacrifices they have made on my behalf. Their prayers for me have kept me going this far.

Soumyabrata Chatterjee
Master of Engineering
Roll No.: 002010502029
Registration No.: 154153 of 2020-21
Department of Computer Science and Engineering
Jadavpur University, Kolkata

ABSTRACT

Bengali language is a rich agglutinative language having compound letters as well as compound words. Compounding is one of the most common method of new word formation in Bengali in which new words are generated by combining two (rarely more) root words. In machine translation or information Retrieval, context of a word is more important to understand its use and meaning. For this purpose, root words play a major role. It is noted that in Bengali new compound words are very often generated by combining two or more words or stems following the word-formation rules and methods applicable in the language to satisfy the linguistic needs of the language. For some of the words, the compound word needs to be split as it is morphologically difficult to analyse it and if not split, may degrade the performance of NLP applications. Sandhi splitting is an important step in NLP applications for languages having compound words formed by Sandhi rules. Here a statistical sandhi splitter for Bengali compound words is proposed. Our approach uses Conditional Random Field (CRF) which is one of the most successful statistical learning methods in NLP for labelling and segmenting sequential data. CRF is trained to find the splitting point of compound words where actual morphological changes occur. From the segments obtained after splitting, the CRF model is again used to find the class label or the sandhi rule that was applied for the formation of the given compound word. 515 words from standard Bengali text book is taken to prepare the dataset. Using this split point and predicted label the root words of the given compound word is determined. Previous tasks for compound splitting mainly were based on rule-based approach and used vocabulary to determine root words. Here our proposed model can determine out of vocabulary words as well and is faster and requires less manual effort. The model could achieve an accuracy of 90% for segmentation stage and 83% for label assignment stage.

CONTENT

List of Figures

List of Tables

List Of Abbreviations

NLP: Natural Language Processing

CRF: Conditional Random Fields

MT: Machine Translation

MDL: Minimum Description Length

LSTM: Long short-term memory

BLEU: Bilingual evaluation understudy

Chapter 1
INTRODUCTION

When two or more words combine to form a new word, that behaves as a single word and usually has different meaning from its constituent words, it is called compound word. A compound word may be defined as a word that has more than one stem word. The process of compound word formation and their functional types may differ from language to language. The process of compounding is done by combining more than one stem word either in its free form or in bound form, which is then integrated lexically to function as single lexical units. The main reason for formation of compounds is to use shorten form for the constituent expanded words. The advantage of compounding lies in making a lengthy sentence shorter and more compact Compound words increase the vocabulary and make the language richer. Presence of compound word may lead to certain problems in natural language processing applications like machine translation, information retrieval, speech recognition, text classification and others. For some of the words, the compound word needs to be split as it is morphologically difficult to analyse and if not split, degrades the performance of NLP applications. There are various methods of compound word formation and Sandhi is one of them. Sandhi splitting acts as an important primary step in NLP applications for agglutinative languages containing compound words formed by Sandhi rules

1.1 Target language (Bengali)

Bengali (also known as Bangla) is typologically an agglutinative language mainly spoken in the Indian sub-continent. Bengali belongs to the Indo-Aryan family of languages that evolved from Sanskrit language. Bengali is a highly inflectional language having more than 160 different inflected forms for verbs and 36 different forms for nouns, and 24 different forms for pronouns. Bengali is the fifth most-spoken native language and the seventh most spoken language by total number of speakers in the world. It is the national language of Bangladesh (144 million speakers – 98% of total population – ranked first according to Bangladesh census 2001) [1] and official (and regional official) language of the states of West Bengal, Tripura and parts of Assam in India (80million speakers – 8.3% of the total population – ranked second according to Indian census 2001).

There is significant difference in dialect between the spoken language of the Bengalis living on the western side and the eastern side of the Padma River. During standardization of Bangla language during the late 19th and early 20th century, the cultural elite were mostly from West Bengal, especially Kolkata (formerly Calcutta). Hence, the dialect of that area was considered the standard for Bangla language. However, at present, the accepted standard language in West Bengal and almost all parts of Bangladesh are identical, i.e., the West Bengal variety. In addition to this, there is another important division in Bengali language: Shadhu bhasa and Cholit (or cholti or cholito) bhasa. The main differences between Shadhu and Cholit forms of Bangla is the adherence to traditional grammar (i.e., the archaic forms of Medieval

Bengali) and a heavily Sanskritized vocabulary in Shadhu bhasha. However, Shadhu bhasha is not spoken in common places and locality but confined to some literary works, novels and formal contexts. So, sadhu bhasa is the old literally form. Here for this work, we will be using mainly the standard colloquial Bangla, i.e., cholit bhasa and hereafter the term “Bangla” will denote only the colloquial Bangla.

1.2 Compound words in Bengali

The regional languages in India are morphologically much richer than many European regional languages. Bengali language is the mother tongue of a large population in India. It is one of the oldest languages and morphologically much richer than many other languages spoken in different parts like Hindi, Marathi.

Compounding is one of the most fertile processes of new word formation in Bengali in which new words are generated by combining two (rarely more) words (or stems). There are large number of compound letter (juktakkhor like ক্ষ = ক্+ষ, ঞ্জ = ঞ্+জ) and compound words (like প্রতিজ্ঞাবদ্ধ , জগজ্জননী). Compound words are formed in bengali by combining two or more stem words following the word combination rules. The method of compounding may be said to be one of the most versatile method by which various new words can be formed thereby increasing the vocabulary of the language.

For our concerned language, Bengali, there are various method of formation of compound words. In general, four types of compound words are found to be used in various Bengali texts and novels[2]. These are:

(a) Tatsama compounds: These compound words are directly inherited from Sanskrit.

(b) Indigenously formed compounds by way of derivation and sandhi rules.

(c) Compound words borrowed from foreign languages (such as Arabic, Persian, English, and others).

(d) Analogically formed compound words, following the word formation rules allowed in Bengali and using various native and non-native lexical elements.

In our work we are concerned with those compound words that are formed using Sandhi rules.

Usually, Bengali compounds are made up with two constituent words, each one of which belongs to a particular lexical category. Rarely, three or more words may be combined together to generate either a single word unit or multiword unit (e.g., হাজার-হাত-কালী i.e “goddess Kali with a garland of thousand cut-off hands”, etc.), For our proposed work we focus only on words formed by combining only two words.

Some Compound Words in Bengali	
বিদ্যার্জন	অনুমত্যনুসারে
লোকসভা	অগ্ন্যুদগার
হাত-পাখা	রাজপথ

Table 1: Compound words in Bengali

Here from the above table, the words বিদ্যার্জন, অনুমত্যনুসারে, অগ্ন্যুদগার are the compound words formed by applying Sandhi rules. Sometimes compound words may be formed just by adding two words by hyphen '- 'sign (like হাত-পাখা). These words can be split easily just by identifying the hyphen mark and separating the words. For this work we are only concerned with words formed by applying Sandhi rules.

1.3 Morphology

Morphology is defined as the branch of linguistic that deals with the study of structure of words. Morphological analysis is the arrangement and relationship of smallest meaningful unit in a language. The smallest unit of a word that cannot be divided further is known as morpheme. Morpheme can be of two types: free morpheme and bound morpheme. Free morpheme is the morpheme that can occur independently, while bound morpheme must attach itself with free morpheme. Example: The word হাতের, can be divided into two parts: হাত, which is the free morpheme and এর which is the bound morpheme (suffix).

A Bengali word consists of three parts: suffix, prefix and lemma. Lemma is the root word. Suffix is added after the lemma and prefix is added before the lemma. Depending on the presence or absence of prefix and suffix, there can be four types of words:

Prefix + lemma: Example আমরণ , here prefix = আ and lemma = মরণ

Lemma + suffix: Example ছেলেটি , here lemma = ছেলে and suffix = টি

Prefix+ Lemma + Suffix: অপ্রত্যাশিত, here prefix =অ, lemma= প্রত্যাশ , suffix= ইত

Lemma+ Lemma: Example হাত তালি , here lemma = হাত and lemma = তালি

Addition of suffix or prefix may change the part of speech of the word.

1.4 Bengali Sandhi

Sandhi has its origin from Sanskrit word 'samdhi' meaning "combination". It refers to a set of morphophonological changes i.e., fusion of final and initial sounds or characters at either morpheme or word boundaries. Sandhi is the medium of creating words based on the Bengali grammar rules**.** The meaning of sandhi is Milan in Bengali which means two words combine

to create a new word. Sandhi is used to make words sweet and making pronunciation easy. In Bengali sandhi is mainly of three types: - 1) Swar-sandhi, 2) Byanjan Sandhi and 3) Bisorgo Sandhi

Swar sandhi: There are 12 vowels in Bengali (‘অ’ ,’আ’ , ’ই’ ,’ঈ’,’উ’,’ঊ’,’ঋ’, ’ঌ’,’এ’,’ঐ’,’ও’,’ঔ’). Swar sandhi rules are created based on these vowels and their combinations of which vowel combine with which vowel and create a new word. Swar-sandhi compound words are obtained by combining two swar barnas.

Compound word using Sandhi	Split words
সিংহাসন	সিংহ + আসন
রবীন্দ্র	রবি + ইন্দ্র
দুর্গোৎসব	দুর্গা + উৎসব
মহৈরাবত	মহা + ঐরাবত
মহৌষধি	মহা + ওষধি

Table 2: Examples of swar sandhi

Byanjan sandhi: There are 39 byanjan barnas or consonants in Bengali literature. Byanjan barnas with swar barna or swar barna with byanjan barna or byanjan barna with byanjan barnas are combined to make new words based on sandhi rules, called Byanjan sandhi. Byanjan-sandhi compound words are obtained by combining two byanjan barnas or one byanjan barna with one swar barna.

Compound Words using sandhi	Split words
উল্লাস	উদ্ + লাস
বৃক্ষচ্ছায়া	বৃক্ষ + ছায়া
গন্তব্য	গম্ + তব্য
সংবাদ	সম্ + বাদ
উজ্জীবিত	উদ্ + জীবিত

Table 3: Examples of byanjan sandhi

Bisorgo sandhi: Here the compound word is obtained by combining bisorgo borno (ঃ) with either swar barna or byanjan barna.

Compound Words using sandhi	Split words
দুশ্চিন্তা	দুঃ + চিন্তা
নিরাকার	নিঃ + আকার
শিরোধার্য	শিরঃ + ধার্য

Table 4: Example of bisorgo sandhi

For this work we will focus on the Bengali compound words formed by the application of sandhi rules. The main objective of this work is to split these compound words using a statistical method so that its performance for various NLP application increases.

The rest of the thesis is organised as follows:

Chapter 2: This chapter reviews the related works on the topic of splitting compound words, their characteristics, advantages and disadvantages.

Chapter 3: This chapter discusses the methodology followed for the proposed work, models used, workflow of proposed model etc.

Chapter 4: This chapter discusses the results obtained using the proposed model and analyses the results.

Chapter 5: This chapter provides a brief conclusion, discusses about the challenges faced and future scope of the project.

Chapter 2
Literature Survey

Compound words splitting is an important pre-processing step for various NLP applications. Research was carried out in different languages to handle the compound words taking into consideration the various methods of compound word formation with respect to the target language. Many researchers proposed different methods and algorithms to extract the root words from the compound words. In Indian literature compound word problems are encountered in various languages and in all cases, researcher came up with different solutions. Three main methods that are used for splitting compound words formed by Sandhi are rule based systems, statistical systems and hybrid systems.

In [3] Koehn and Knight (2003) proposed an empirical splitting algorithm for machine translation from German to English. It introduces method to learn splitting rules from monolingual and parallel corpora. This method split words in all possible places, and considered a splitting option valid if all its parts had been seen as words in a monolingual corpus. This algorithm allowed the addition of *-es* or *-s* at all splitting points. If more than one valid splitting options are present, then they are selected based on the number of splits, the geometric mean of part frequencies, or based on alignment data. Three different objectives and hence different evaluation metrices are used here: one to one correspondence, word-based translation system and phrase-based translation system. The algorithm was evaluated intrinsically on a gold standard of manually split noun phrases and on machine translation of noun phrases. The best results for machine translation were achieved by using either the geometric mean, or the highest number of splits. According to the authors, one typical error of this proposed method is that prefixes and suffixes are often split off leading to wrong splitting. To tackle this error information about parts of speech of words may be used so that compounds are not split into preposition and determiners. The result show accuracy up to 99.1% and BLEU score of 0.039 for German-English noun phrase machine translation.

In [4], author proposed to improve the performance of statistical machine translation using morphological analysis. In statistical machine translation, the problem of sparse data can make estimating word-to-word alignment probabilities for the translation model difficult. Here the aim was Czech to English machine translation. Czech is a highly inflected language and so the problem of sparse data is more severe. This work showed that using morphological analysis to modify the Czech input can improve a Czech-English machine translation system. Here various methods of incorporating morphological information were evaluated, and it was shown that the system that combines these methods yields the best results.

Morphological variations in Czech are reflected in several different ways in English. In some cases, such as verb past tenses or noun plurals, morphological distinctions found in Czech are also found in English. In other instances, English may use function words to express a meaning that occurs as a morphological variant in Czech. For example, genitive case marking can often be translated as of and instrumental case as by or with. In still other instances,

morphological distinctions made in Czech are either completely absent in English (e.g., gender on common nouns) or are reflected in English syntax (e.g., many case markings). Handling these correspondences between morphology and syntax requires analysis above the lexical level and is therefore beyond the scope of this paper.

This method simply identifies the pseudowords (words which is less frequent) and replace with their lemmas which is identified by their past experience. In general, in statistical MT, we simply mapped one word of a language to one word of another language and if this mapping couldn't be possible then simply use the linguistic rules to generate the correct lemma.

In [5], The paper discusses the techniques and results of developing an algorithm, which accepts raw linguistic data as input and produces as their output an analysis of data or a grammar. The main purpose for the work is to learn morphology on the basis of essentially no prior knowledge save for the data. The basic aim described is the determination of location of the breaks between morphemes inside any word. Here the author reviews Zellig Harris' idea (*From phoneme to morpheme ,1955)* that morpheme boundaries occur at positions of maximum phoneme choice. Based on the shortcomings of this idea, author John Goldsmith proposes to divide the process of morphological analysis into 'a set of heuristics' and 'Minimum Description Length (MDL)' evaluation process. Linguistica is based on the MDL principle, which states that the optimal hypothesis to explain a set of data is the one that minimizes the total number of bits required to describe both the hypothesis and the data under that hypothesis. The heuristics then is divided into 'initial bootstrapping heuristics' which determines the first analysis of stems and suffixes, and 'incremental heuristics' which modifies this analysis. The MDL then decides whether the modifications made by the incremental heuristics should be adopted or dropped.

In [6], a rule-based learning method for root words in Hindi language was proposed. In this paper inflectional lemmatizer was created which generates the rules for extracting the suffixes and also added rules for generating a proper meaningful root word. For the method proposed, the first step is to read the input word. A database is created containing root words along with features like gender, numbers etc. The database contains all the root words. This input word is checked in the database and if it is present in the database then the given input word is displayed as it is. If the word is not present in the database, then it comes down to access the rules. After accessing the rules, the root word is generated using lemmatizer and displayed. The root word generated is added to the database and vocabulary is enriched. Typically, lemmatizer is built using a rule-based approach and paradigm approach. For rule-based approach along with the rules, knowledge base is created for storing the grammatical features. This knowledgebase creation requires a large amount of memory, but with respect to time it can produce the best, accurate and fast result. This method aimed at time complexity optimization. The main drawback of this proposed method is that it mainly focused on nouns, the adjective and verb they didn't think about broader prospects. Here lemmatizer is used to lemmatize the words which is not a good approach as it modifies the previous original input words.

The paper [7] proposed a modified method using sequence-to-sequence model used in morphological analysis of Sanskrit. Sanskrit poses considerable challenges for NLP at the levels of tokenization, lemmatization, and morphological analysis mainly due to compound

words formed by sandhi rules. This method uses two step approaches. In the first stage sandhis are resolved and most probable lexical reading is detected using a bigram model. In the next step using rule-based morphological analyzer.in this first a tag data set was created. The inflection morphology of Sanskrit is mainly described in five categories noun, pronoun, verb, adjective, and verbal particles. To reduce the Collison, tag sets are created.
According to this paper, 86 tags are created depending on various factors. After this data classification was done. To check the performance two type of classifiers are used one is CRF (Conditional Random Fields) based classifier, another is sequence-sequence model-based classifier. In sequence-to-sequence model mainly three layers of LSTM units are used as follows:
1. A fully connected input layer with an embedding size of 70, tanh activation, and a subsequent dropout layer with a dropout rate of 20%

2. Two bidirectional LSTM units.

3. An output layer that's fully connected to the output of the second bidirectional LSTM.

According to this work, using a reduced tag set, a combination of morphological, lexical, and semantic features, and a bidirectional deep neural network, the accuracy rates for morphological disambiguation could be improved significantly in comparison to previously published results.

In [8] the author developed a fully statistical sandhi splitter. The model was tested with Dravidian languages like Telegu and Malayalam, which are rich agglutinative language and it gives accuracy up to 90.5%. This method uses two stages to split compound words and find their root words. Both the stages use CRF to handle sequential data.

This model is language independent mainly because of automatically extracted class labels. Even though the given model handled sandhi in one type of compound words, this model can be readily adapted to other types as well.

The paper [9] extends the work proposed in paper [8] for various NLP applications. This paper discusses in detail the analysis of Statistical Sandhi Splitter for three main areas of NLP namely: machine translation, anaphora resolution and dialogue system. It was observed that the presence of compound words degrades the performance of any NLP applications and this can be improved significantly using the proposed model of statistical sandhi splitter.
The splitting of some compound words depends on the context in which the words are used. These words may be split incorrectly using this method. So, in order to eliminate this error, the entire sentence may be considered during training instead of just the words in order to include the contextual information.

Sandhi splitting tasks for various regional languages have been studied. The sandhi splitting tasks mainly was carried out using three methods: Rule based, Statistical system, and hybrid system.

Type	Characteristics
Rule based System	Building rule-based systems to identify different words in the compound word. It was built for Malayalam and Marathi language. Drawback: they require a lot of manual effort and time to prepare rules
Statistical System	Simple finite state automata was used for finding possible words in a given compound word. Drawback: This approach fails for out-of-vocabulary words i.e. if base word of any compound word doesn't exist in vocabulary.
Hybrid System	Combine both statistical and rule-based techniques. It was tested for Malayalam language. It identifies split point statistically and uses character level rules specific to language to split the compound word accordingly.

Table 4: Related work on sandhi splitting

In 2015, Kuncham[8] proposed a statistical model for splitting sandhi words in Telegu and Malayalam language. Taking motivation from that work, here we proposed statistical sandhi splitter for Bengali words.

Our concerned language Bengali is rich in sandhi-based compound words. There has not been much work related to sandhi splitting in Bengali language. Some rule-based methods were proposed earlier which requires a lot of manual work. Here our proposed system is based on statistical method using Conditional Random fields. On comparing with traditional rule based or hybrid methods, it is found to be faster, more robust and require less manual effort.

Chapter 3
Proposed Work

Agglutinative languages contain compound word in which two or more different root words are combined. Bengali is a rich agglutinative language containing a lot of compound words. Sandhi is one of the methods by which these compound words are formed. The objective of our work is to split compound words in Bengali language to its constituent words. Sandhi Splitting would act as an important pre-processing step for NLP applications in Bengali language.

Here we will be using Statistical approach for the task of splitting compound words. The splitting points would be determined statistically from the knowledge of the training words. After obtaining the splitting points, the actual words are formed by changing the word boundaries following various Sandhi rules.

For Sandhi compound words, the sandhi rule to be applied for formation of compound words depends on the letters at the word boundaries. So here, the contextual information and state of neighbouring letters are crucial in determining the sandhi rules. For this purpose, conditional random field (CRF) is best suited model. Unlike other traditional classifier that predicts the label for single sample without considering its neighbours, the CRF takes the state of its neighbour into context for predicting the labels.

3.1 Model Used (Conditional Random Field – CRF)

Conditional Random Field (CRF) are statistical modelling method applied in various pattern recognition and machine learning tasks like part of speech tagging, labelling and parsing of sequential data, named entity recognition, gene prediction etc. where contextual information affects the current prediction task. Conditional Random field is a class of discriminative model where the state of neighbour is taken into account for prediction of labels.

CRFs can be regarded as a type of probabilistic graphical model. In CRF the predictions are modelled as graphical model, which represents presence of dependencies in prediction.

CRF on observations X and random variables Y is defined as:

Let G = (V, E) be a graph such that Y= (Y_v) v € V, where the vertices of graph V represent Y

Then conditional random field is (X,Y) where each random variable Y_v , conditioned on X obeys Markov property with respect to graph, that is the probability is dependent on the neighbours in G.

$P(Y_v \mid X, \{ Y_w : w \neq v \}) = P(Y_v \mid X, \{ Y_w : w \text{ is neighbour of } v \})$

So, in CRF model the nodes can be divided into two disjoint sets X and Y, where X is observed variable and Y is the output variable and then the conditional probability P (X | Y) is then modelled depending on whether the output variables are neighbours or not.

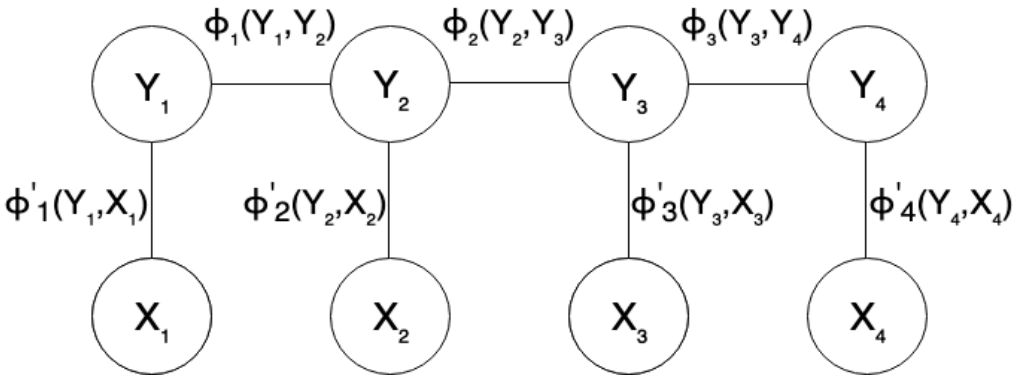


Figure1: Conditional Random Field structure

Since CRF is a discriminative model, that is it models the conditional probability P(Y|X) i.e. X is always given or observed. Therefore, the graph ultimately reduces to a simple chain.

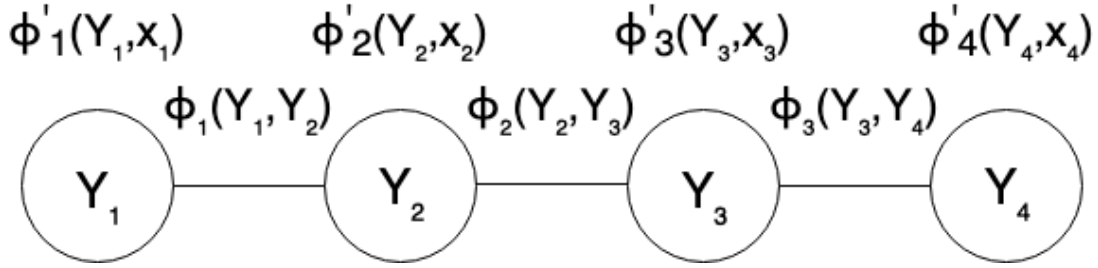


Figure 2: CRF model conditioned on X

Conditioned on X, we will be trying to find the corresponding Y_i for every X_i

3.2 Methodology

This proposed statistical system for sandhi splitting works in mainly two major stages: segmentation and word generation.

3.2.1 Segmentation: Segmentation is the first stage in splitting of compound words. Segmentation, as the name suggests means splitting or breaking down. The input for this stage is the compound words and the output is the segments that show boundary/split points in the given input word.

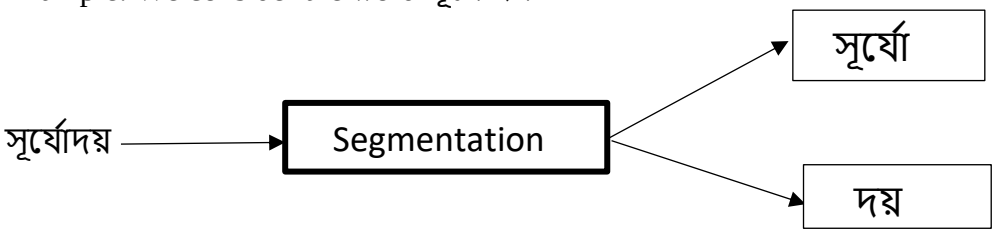
In this stage for each character in the compound word the CRF model decides whether to split or not split the compound word at that point. At each character the CRF determines the probability of splitting the word at that point. Here we assume each word consists of two segments, so the probability of splitting at each character is determined using CRF model and the character giving maximum probability of splitting is taken as the split point.

Most of the previous methods relies on vocabulary for splitting a word. At each point it determines whether we get a meaningful word if split at that point, but it will not always give correct result for sandhi compound words as the segments are not always meaningful words in this case. So, using this statistical method, the performance can be improved as well as it can predict split point of words not in vocabulary list.

Thus, this stage identifies the word boundaries where actual morphological change occurs. The segments obtained as output may or may not be meaningful always.

The CRF model can be trained using feature set like characters and character tags for this segmentation stage. The actual morphophonological changes occur in characters so this feature is used to identify the exact location where these changes occur. Character tags are assigned based on the sound of the character and is used to incorporate the information regarding types of vowel/consonant information that occur during morphophonological changes. For Bengali words character tag may be like swor barna(vowel) or byanjan barna(consonant).

Example: We consider the word সূর্যোদয়



For segmentation stage: input = সূর্যোদয়
Output = সূর্যো and দয়

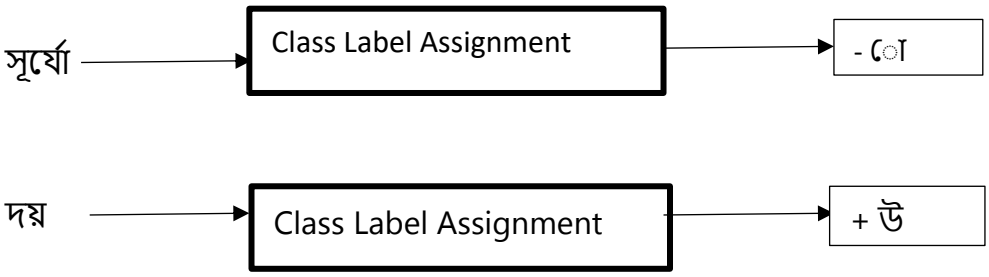
For the subsequent stages as well, we will use the same example to show the working of the proposed method clearly.

3.2.2 Word Generation:

The segmentation stage is followed by the word generation stage. As we have discussed earlier, the segments obtained in the segmentation stage may or may not be meaningful words. So, in this stage from the segments generated in the previous stage, the meaningful words are generated using the rules of sandhi applicable. The input to this stage is the segments of the compound word obtained in previous stage and the outputs are the meaningful words. The main challenge for this stage is to determine which sandhi rule has been applied for the compound word formation, so that the correct root word can be generated.

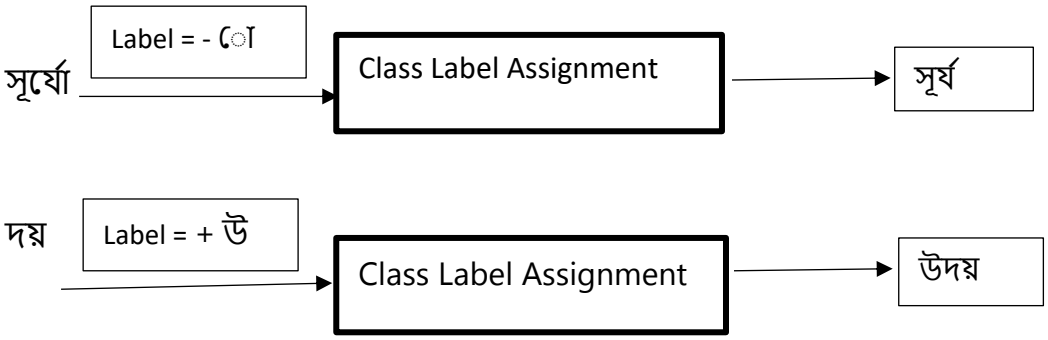
This word generation stage can be further divided into two stages namely: class label assignment and word formation stage.

Class label assignment: In class label assignment stage, the word segments are assigned a class according to morphological changes of addition or deletion of characters in Sandhi.
Changes in character occurs at the word boundaries, that is for the first segment there is change at the last few characters and for the second segment there is change in first few characters, if any. So, here the CRF model can be trained using feature set like segments and prefix and suffix characters.
For this stage, the input is word segments and output is class label showing addition or deletion of character.
Example: Continuing from the previous stage of segmentation for the word সূর্যোদয়,



For the first segment, input = সূর্যো, output label = - ো
For the second segment, input = দয়, output label = + উ
In the output label '+' means addition of the character and '– 'means deletion of the character.

Word formation: In the last and final stage, using the information from the class labels meaningful words are generated from the segments. The input for this stage is word segments and class label. Using the class label output generated is meaningful root words.
Example: The segments and corresponding class labels were:
1. সূর্যো , label = ো
2. দয়, label= + উ



So, the final output are সূর্য and উদয়, which are meaningful Bengali words obtained from Sandhi compound word সূর্যোদয় .

সূর্যোদয় (sunrise)= সূর্য (sun) + উদয় (rise)

The workflow of proposed methodology is given below:

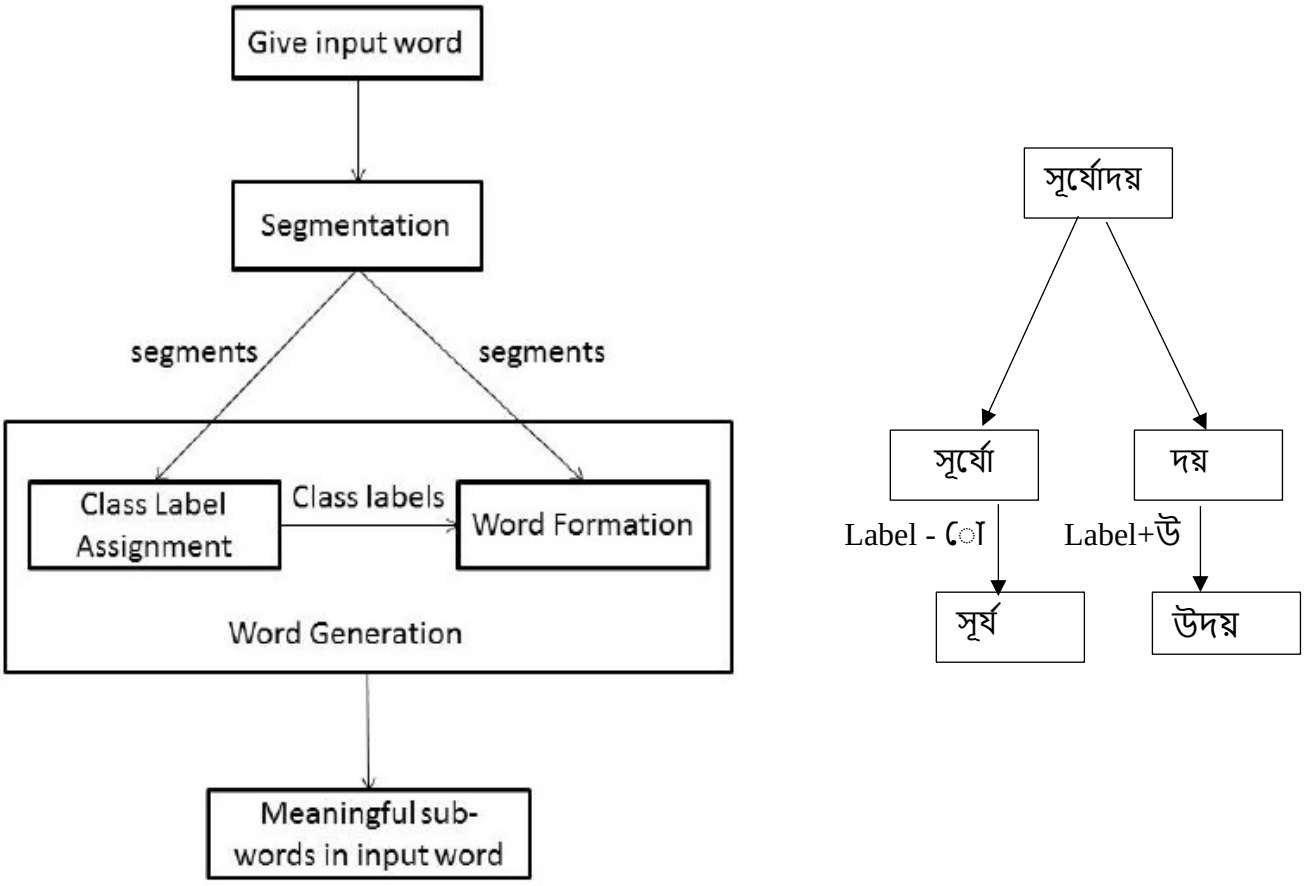


Figure3: Flow chart of Sandhi Splitter

Chapter 4
Result and Analysis

In this section the dataset used for the model and the results obtained are discussed.

4.1 Dataset

For training the CRF model we require annotated dataset containing Sandhi compound words and their corresponding split words. But this annotated dataset for Bengali language was not available. Monolingual corpus from TDIL [10] dataset contains Bengali text corpus from various fields were available but not exactly suitable for the task. So, the dataset was prepared taking sandhi compound words and their split words from standard Bengali grammar text books[11]. This dataset would act as gold standard data.

515 compound words along with their split root words was used as dataset for training the model using CRF. Out of these 515 words, 230 words are swar sandhi, 182 words are byanjan sandhi and remaining 103 words are bisorgo sandhi compound words. The same dataset is used for both segmentation and class label assignment tasks.

Type	Number of samples
Swar-sandhi	230
Byanjan-sandhi	182
Bisorgo-sandhi	103
Total	515

Table 6: Sandi distribution in dataset

In the first stage CRF model extract features from the Sandhi compound words and uses the information from the neighbouring characters to predict the splitting point of the words. In the next stage the CRF model again extract features from the split segments and then predicts the class label of the split words.

4.2 Templates used and result:

As discussed, the proposed model splits the compound words in two stages. In both these stages CRF is used for prediction. CRF is used as it can handle the contextual information from its neighbour as well during prediction of label or prediction of split point for compound words.

For CRF model a proper feature set has to be chosen for accurate result from the model. For the segmentation task, at each character we need to decide whether to split at that point or

not. So here we need characters as feature. For each character, different number of characters on left and right of current character is included as feature and it is seen 4 characters to the left and 4 characters to the right of current character gives the best result.

To get the mean accuracy cross validation technique was carried out, so that the entire data set can be used for training and testing purpose and the it can accurately estimate how the model will perform.

Template	No of characters to the left and right	Accuracy (%)
1.	3	90.125
2.	4	90.60
3.	5	90.21

Table 7: Templates for segmentation

For each character 4 characters to the left and 4 characters to the right of current character is taken as its feature, giving an accuracy of 90.6% for splitting. Along with each character whether it is swar barna(vowel) or byanjan barna (consonant) is also taken in the feature set.

Example: Let the compound word be হিমালয় and current character be ‘ম’ , then the feature set with 4 character to the left and right of current character will be :

{'+1:char': 'া', '+1:chartag': 'sb', '+1char|char': 'মা', '+2:char': 'ল', '+2:chartag': 'bb', '+3:char': 'য়', '+3:chartag': 'bb', '-1:char': 'ি', '-1:chartag': 'sb', '-1char|char': 'িম', '-2:char': 'হ', '-2:chartag': 'bb', 'char': 'ম'}

Here ‘- ‘represent characters to the left of current character and ‘+’ represent characters to the right of current character.

The output of the segmentation will be the index with maximum probability of splitting. The assumption taken here is that each word consists of only two segments, so the output produced will be a single index for splitting the word.

For the next stage, that is class assignment stage, the split segments will be taken as input for the CRF model. Features will be extracted from these segments and will be used to predict the label or Sandhi rue for the particular compound word formation. The features used for this stage are prefix and suffix character. For sandhi-based compound words, changes take place in the suffix part of the first segment and prefix part of the second segment. (Ex: w1x + yw2 = w1zw2). The other portion of the word segments except the last part of first segment and first part of second segment are not involved in the splitting and remains unchanged. So, they need not be considered for feature set in the label assignment task involved. The feature set for the label assignment stage is suffix of first segment and prefix of second segment.

Different number of characters are taken as prefix and suffix to obtain the best feature set and the accuracy is calculated using cross validation method.

Template	No of characters taken as suffix (for 1st segment)	No of characters taken as prefix (for 2nd segment)	Accuracy (%)
1.	2	2	82.2
2.	3	3	83.78
3.	4	4	83.1

Table 8: Template for label assignment

Taking 3 character as prefix and suffix gives better result with accuracy of 83.5 %.

Example: For the split segments ‘হিম’ and 'ালয়', the feature set would be:

{'cur-1': 'ম', 'cur-2': 'িম', 'cur|nex': 'মা', 'next1': 'া', 'next2': 'াল', 'word': 'হিম'}

and

{'cur1': 'া', 'cur2': 'াল', 'prev-1': 'ম', 'prev-2': 'িম', 'prev|cur': 'মা', 'word': 'ালয়'}

For training the CRF at this stage 515 Bengali words were used, that is, 1030 word-segments were used to train the model. From the dataset 122 different labels (sandhi rules used for formation of compound words) were extracted.

The final stage just modifies the word segments using labels obtained. The characters followed by ‘- ‘are removed from the segment and characters followed by ‘+’ are added to the segment. Again, since any modification occurs at the word boundaries, so any addition or deletion are carried out at the ending of first segment and the beginning of second segment.

Name of the stage	Accuracy (%)
Segmentation	90.6
Class Label assignment	83.78

Table 9: Accuracy of different stages

The final stage, modification occurs at the word boundaries according to the predicted label. So here accuracy calculation is not required. Accuracy of word formation and word generation stage is same as accuracy of class label assignment stage.

4.3 System requirements

Python version 3.7.13 was used for the proposed work. Google collab notebooks were used that run in the cloud and are highly integrated with google drive for easy access and sharing of resources and faster implementation.

Minimum RAM requirement = 1GB, Operating system Windows 10 was used.

For the CRF module sklearn-crfsuite was used which provides a scikit-learn compatible estimator. scikit-learn version 0.22.2 was used to meet the dependencies. CRF API reference: Class sklearn_crfsuite.CRF. It is a python-crfsuite wrapper with interface siimlar to scikit-learn. It allows to use a familiar fit/predict interface and scikit-learn model selection utilities (cross-validation, hyperparameter optimization).

Chapter 5

Conclusion

Here we have proposed a statistical sandhi splitter for Bengali compound words. It was observed that presence of compound words acts as a barrier for NLP applications and degrades the performance. So, to improve these tasks we have proposed sandhi splitting as a pre-processing step. The focus language is Bangla here, which is a rich agglutinative language having high percentage of compound words in the vocabulary. For the compound words formed from Sandhi rules, not necessarily both the constituents are valid dictionary words. So, frequency-based approaches which searches for valid dictionary words does not give satisfactory results. Here lies the significance of this proposed approach.
Compound word formation by application of Sandhi rules is a relatively easier task, but the reverse process of decompounding, that is splitting the compound words is comparatively difficult task. Here we need two-fold prediction task, one for finding out the split point and another to predict the particular sandhi rule that had been applied to form the given word. Alongside this, the accuracy of the entire model depends on the accuracy of the sub stages, so if the substages has lesser accuracy, then the subsequent stages and hence the entire model will have less accuracy.

5.1 Challenges

The main challenge for the NLP task in Bengali is non availability of dataset containing annotated data of compound words along with their meaningful split words. For this task gold standard dataset of 515 was prepared from standard text book. The accuracy can presumably be increased further if more data is available for training the model.
Another challenge was that for some words the word has different meaning compared to its constituent words. These words if split may degrade the performance for NLP applications.
For example: লোকসভা means the lower house of the parliament, but its constituent root words লোক means people and সভা means meeting. So, the compound word has different meaning from its constituent root words. Thus, a decision has to be made to determine whether the word needs to be split or not. Some words depend on the contextual information to be split and for these we need to consider the entire sentence instead of just words for training. Some words also have dialectal influence which is not handled in this work. So, we may extend this project to consider contextual information and non-standard language in future.
For Sandhi splitting task another major challenge was the presence of নিপাতন-সন্ধি which are sandhi compound words but do not follow the sandhi rules. (Example: হিন্স্ + অ = সিংহ).
Nipatan Sandhi is present in Swar sandhi as well as byanjan sandhi rules. These act as an exception to sandhi rules These words pose a challenge in the class assignment as they will not follow the traditional rules and hence this will lead to wrong split for most of the words. Presence of these compound words decreases the accuracy of the model and hence degrades the performance.

5.2 Future scope

This work may be extended adding an extra stage to detect compound words from Bengali text corpus and then split those detected words. Here we are providing the compound words and then splitting those words, if the model is trained using sandhi words and non-sandhi words so that it can predict those words that need to be split then it can automatically detect the compound words and then split only those words that are needed to split, leaving the other words as it is. We can integrate it with our proposed model to automatically detect Bengali compound words and split them from text corpus.
Here we have prepared standard dataset, of 515 words. Further we may increase the dataset to a greater number of words and compare the accuracy of the model.

We can further show that the splitting of words actually increases the performance of various NLP tasks like anaphora resolution, dialogue system etc. For this we will need to feed the same dataset without splitting and with splitting the compound words, then we need to compare the performance for each task and check if accuracy increases on splitting the compound words.

As already discussed, some words need to be split depending on contextual information, so considering the entire sentence to determine whether to split or not split the words and taking the regional dialectal influence into consideration would be part of our future work.

Reference

[1] *"5 Surprising Reasons the Bengali Language Is Important". 17 August 2017. Archived from the original on 26 June 2018. Retrieved 10 March 2018*

[2] *Dash, Niladri. (2020). The Morphodynamics of Genitive Case Markers in the Formation of Inflected Words in Bangla: An Empirical Study on the Bangla Lexical Database. 4. 26-40.*

[3] Koehn, P., & Knight, K. (2003). *Empirical methods for compound splitting. Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - EACL '03.*

[4] Improving Statistical MT through Morphological Analysis, October 2005, Conference: HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada

[5] Goldsmith, John. (2000). Linguistica: An Automatic Morphological Analyzer.

[6] Design of a Rule Based Hindi Lemmatizer, Snigdha Paul, Mini Tandon, Nisheeth Joshi and Iti Mathur, July 2013, Third International Conference on Advances in Computing & Information Technology

[7] Improving the Morphological Analysis of Classical Sanskrit, A Hellwig, Oliver, Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)2016, 8 Dec, The COLING 2016 Organizing Committee, Osaka, Japan

[8] Kuncham, P., Nelakuditi, K., Nallani, S., and Mamidi, R. (2015). Statistical sandhi splitter for agglutinative languages. In Computational Linguistics and Intelligent Text Processing pages 164–172, Springer.

[9] Kuncham, Prathyusha et al. "Statistical Sandhi Splitter and its Effect on NLP Applications." *RANLP* (2015).

[10] The TDIL Program and the Indian Language Corpora Initiative (ILCI).
January 2010
Conference: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta

[11] **উচ্চতর বাংলা ব্যাকরণ** শ্রী বামনদেব চক্রবর্তীর, revised 8th edition, march 2013