Morphological Stemming Cluster Identification for Bangla

Amitava Das

Jadavpur University

Department of Computer Science and Engineering,
Jadavpur University, Kolkata 700032, India

amitava.santu@gmail.com

Sivaji Bandyopadhyay

Jadavpur University

Department of Computer Science and Engineering,
Jadavpur University, Kolkata 700032, India

sivaji_cse_ju@yahoo.com

Morphological Stemming Cluster Identification for Bangla

Abstract

This paper describes Morphological Parsing of Bangla (ethnonym: Bangla; exonym: Bengali) words using stemming cluster technique. The addition of inflectional suffix; derivational suffix and agglutination in compound words make Morphological Parsing fairly complex for the Bangla. There are existing efforts at building a complete morphological parser for Bangla [1], required for various NLP applications. Morphological Parsing in Information Retrieval aspect does not demand full Morphological feature structure always; rather identification of stems from several surface form of a particular word is required. A Morphological stemmer based on stemming cluster technique has been developed. This feature analyzes prefixes and suffixes of all the word forms present in a particular document. Words that are identified to have same root form are grouped in a cluster with the identified root word as cluster centre. An inflectional suffix is a terminal affix that does not change the word-class (parts of speech) of the root during concatenation; it is added to maintain the syntactic environment of the root in Bangla. On the other hand, derivational suffixes change word-class (parts of speech) and the orthographic-form of the root word. Experiments have been carried out with two types of algorithms: simple suffix stripping algorithm and score based stemming cluster identification algorithm. The Suffix stripping algorithm simply checks if any word has any suffixes (one or more than one suffixes) from a manually generated suffix list and then the word is assigned to the appropriate cluster where cluster centre is the assumed root word, i.e., the form obtained after deleting the suffix from the surface form. Suffix stripping algorithm works well for Noun, Adjective, Adverb categories. The words of other part of speech categories especially Verbs follow derivational morphology. The score based stemming technique has been designed to resolve the stem for inflected word forms. The technique uses Minimum Edit Distance method [2], well known for spelling error detection, to measure the cost of classifying every word being in a particular class. Score based technique considers two standard operations of Minimum Edit Distance, i.e., insertion and deletion. The consideration range of insertion and deletion for the present task is maximum three characters. The idea is that the present word matches an existing cluster centre after insertion and/or deletion of maximum three characters. The present word will be assigned to the cluster that can be reached with minimum number of insertion and/or deletion. This is an iterative clustering mechanism for assigning each word into a cluster. A separate list of verb inflections (only 50 entries; manually edited) has been maintained to validate the result of the score based technique. The standard K-means Clustering technique has been used here. Each cluster centre is treated as a root stem. The system reported an accuracy of 74.6%.

1. Introduction

Normal morphological parsing strategy decomposes a word into morphemes given lexicon list, proper lexicon order and different spelling change rules. But this is not enough to compute the part of speech of a derivationally complex word or return a word's inflectional features. In this paper we will discuss a Morphological stemmer based on stemming cluster technique. Existing effort in literature for Bengali is very less in number.

Sajib Dasgupta and Mumit Khan reported a Morphological Parser for Bangla using PC-KIMMO, which is widely used by linguistics around the world for morphological parsing and generation. PC-KIMMO is based on Kimmo Koskenniemi's famous model of Two-level Morphology in which a word is represented as a correspondence between its lexical level form and its surface level form. In this paper they presented to incorporate Bangla in PC-KIMMO.

Sandipan Sarkar and Sivaji Bandyopadhyay[3] in 2009 presented a full-phrased rule-based stemming for Bengali. The paper presented a detail analysis of corpus and grammar.

Morphological Parsing in Information Retrieval aspect does not demand full Morphological feature structure always; rather identification of stems from several surface form of a particular word is required. A Morphological stemmer based on stemming cluster technique has been developed.

Stemming: the problem in Bengali

Bengali is one of the most morphologically rich languages. Before start any experience we look into formal grammar and previous works. Statistics shows the difficulties and various characteristics of inflections in Bengali.

More than one inflection can be applied to the stem to form the word type. A thorough analysis of NEWS corpus reveals that up to three inflections applied to a stem. Categorically different POS take different number of inflections after stem. Table 1 present the inflections taken by different POS. The count of inflections at n^{th} position after stem, which is designated as P_n is also calculated.

POS	Total	P_1	P_2	P_3
Noun	34	34	6	1
Pronoun	37	34	10	2
Adjective	18	16	3	0
Adverb	8	8	3	0
Verb	129	127	4	1
Conjunction	3	3	0	0
Postposition	5	5	1	0

Table 1: Inflection Counts

But stemming is easier for closed POS types by dictionary based approaches or by other standard techniques. Stemming is a hard problem for the four open POS categories; Noun, Adjective, Adverb and Verb.

But there are categorical differences between of stemming among these four classes. Most general approach to solve the stemming as a problem is as follows;

1. **Lexicon**

   The list of stems and affixes, together with basic information LEXICON about them (whether a stem is a Noun stem or a Verb stem, etc).

2. **Morphotactics**

   The model of morpheme ordering that explains which classes of morphemes can follow other classes of morphemes inside a word. For example, the rule that the Bengali Tense Person Affixes follow the Verbs rather than preceding it.

3. **Orthographic Rules**

   These spelling rules are used to model the changes that occur in a word, usually when two morphemes combine. For example root word hAt (হাট) is changed into hEt (হেট) when added with verb suffix to form a word hEtECI (হেটেছি)

It is very clear from the previous statistics that Verb is the most problematic area for Stemming. Bangla has a vast inflectional system; the number of inflected and derivational forms of a certain lexicon is huge. For example there are nearly (10*5) forms for a certain verb word in Bengali as there is 10 tenses and 5 persons and a root verb changes its form according to tense and person. For example here are 20 forms of verb root KA (খা).

The addition of inflectional suffix; derivational suffix and agglutination in compound words make Morphological Parsing fairly complex for the Bangla. There are existing efforts at building a complete morphological parser for Bangla [1], required for various NLP applications. Morphological Parsing in Information Retrieval aspect does not demand full morphological feature structure always; rather identification of stems from several surface form of a particular word is required. A Morphological stemmer based on stemming cluster technique has been developed.

2. Morphological Clustering

Two types of morphological clustering strategy used here. First one is for agglutinative suffix stripping. A manually generated suffix list has been generated for the present task. The list is sorted according to the length of the suffixes. The second method work with minimum edit distance methodology with a suffix list.

2.1 Corpus-based acquisition of suffix list

The suffix list used here generated semi-automatically from corpus. Part of speech wise four lists have been prepared of Noun, Adjective, Adverb and Verb. A basic clustering technique with threshold value -3 (deletion of three character at the end of the word) to +3 (insertion of three character at the end of the word) has been considered to make clusters of words in corpus. Project Indian Languages to Indian Languages Machine Translation System's (IL-ILMT) shallow parser[1] has been used here.

[1] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

Every cluster centre then considered as a root form of that cluster. A list of suffixes deuced from the other surface forms of that word by subtracting the root word from the surface words. Automatically generated suffix list then sorted out with unique value and manually checked to build up the final list. Table 1 representing a snapshot of the categorically prepare suffix list.

Type	Root	Surface Form	Suffixes
Noun	ভারত	ভারতে, ভারতের	ে, ের
Adjective	অমানব, দুর্ভাগ্য	অমানবিক, দুর্ভাগ্যবশত	িক বশত
Adverb	ভারী, দূর, দূর	ভারিক্কি, দূরীভূত	িক্কি, ীভূত
Verb	খা	খাচ্ছেন, খেয়েছিলেন	চ্ছেন, য়েছিলেন

Table 2: Semi-Automatically Generated Suffix List

2.2 Simple Suffix Stripping

Simple suffix stripping algorithm works well for Noun, Adverb and Adjective classes. It checks every unassigned word with every cluster centre by subtracting suitable suffix from categorical suffix list. The algorithm starts iteration from *k* number of clusters. Where *k* is the total number of word forms present in a particular document. It ended up its iteration with *n* number of clusters. Where *n* is lesser than *k.*

2.3 Clustering

Several words in a sentence that carry opinion information may be present in a sentence in their inflected forms. Stemming is necessary for such inflected words before they can be searched in appropriate lists. Due to non availability of good stemmers in Indian languages especially in Bengali, a stemmer based on stemming cluster technique has been developed. This feature analyzes prefixes and suffixes of all the word forms present in a particular document. Words that are identified to have same root form are grouped in a finite number of clusters with the identified root word as cluster center. The term prefix/suffix is a sequence of first/last few characters of a word, which may not be linguistically meaningful. The use of prefix/suffix information works well for highly inflected languages like the Indian languages. Experiments are carried out with two types of algorithms: simple suffix stripping algorithm and score base stemming cluster identification algorithm. A small list of 205 suffixes for Bengali has been manually generated. The Suffix stripping algorithm simply checks if any word has any suffixes (one or more than one suffixes) from the list and then the word is assigned to the appropriate cluster where cluster center is the assumed root word, i.e., the form obtained after deleting the suffix from the surface form. Suffix stripping algorithm works well for Noun, Adjective, Adverb categories. In case of Verbs in Bengali, root form of the word changes when suffixes are added. Hence for the Bengali Verb words simple suffix stripping does not work well. The score based stemming technique has been designed to resolve the stem for inflected Verb words. The technique uses Minimum Edit Distance method [35], well known for spelling error detection, to measure the cost of classifying every word being in a particular class. Score based technique considers two standard operations of Minimum Edit Distance, i.e., insertion and deletion.

n	9	10	11	10	11	12	11	10	9	**8**
o	8	9	10	9	10	11	10	9	**8**	9
i	7	8	9	8	9	10	9	**8**	9	10
t	6	7	8	7	8	9	**8**	9	10	11
n	5	6	7	6	7	**8**	9	10	11	12
e	4	5	6	**5**	**6**	7	8	9	10	11
t	3	4	**5**	6	7	8	9	10	11	12
n	2	3	**4**	5	6	7	8	8	10	11
i	1	**2**	3	4	5	6	7	8	9	10
#	**0**	1	2	3	4	5	6	7	8	9
	#	e	x	e	c	u	t	i	o	n

Table 3: Computation of minimum edit distance between intention and execution via distance with cost of 1 for insertions or deletions, 2 for substitutions. Substitution of a character for itself has a cost of 0.

The consideration range of insertion and deletion for the present task is maximum three characters. The idea is that the present word matches an existing cluster centre after insertion and/or deletion of maximum three characters. The present word will be assigned to the cluster that can be reached with minimum number of insertion and/or deletion. This is an iterative clustering mechanism for assigning each word into a cluster. The system iterates 6 times i.e. it starts from -3 (deletion of three characters) and ended with +3 (insertion of three characters) value and finally generate a finite number of stemming clusters. A separate list of verb inflections (only 50 entries) has been maintained to validate the result of the score based technique. The standard K-means Clustering technique has been used here. K-means is a hard clustering algorithm that defines clusters by the center of mass of their members. K-means need a set of initial cluster centers in the beginning. Then it goes through several iterations of assigning each object to the cluster whose center is closest.

Stemming Clusters
সত্যজিৎ,সত্যজিৎকে,সত্যজিত,সত্যজিতের
ছবি,ছবির,ছবিটি,ছবিতে,ছবিতেই,ছবিটিতে,ছবিটির
রায়,রায়ের
ওপর,ওপরেও
করত,করতেন
নিয়ে,নিয়েছেন
দেন,দেননি
নির্মিত

Table 4: Bold words are cluster centre.

After all objects have been assigned, a re-computation has been done for the centers of each cluster as the centroid or mean of its member. The manually edited list of suffixes has been re-used here to validate cluster members. Since the manually edited suffix list is not an exhaustive list, the words whose distance from the cluster centre is less than or equal to two characters, are kept in the cluster,

otherwise a separate cluster is created with the word. No evaluation has been done yet for the proposed technique. Each cluster center is treated as a root stem. For English, standard Porter Stemmer algorithm has been used.

3. Evaluation

Evaluation of the present system has been done on Project Indian Languages to Indian Languages Machine Translation System's (IL-ILMT)[2] gold standard Morphological dataset. The dataset has been prepared manually. The dataset consist of 1000 sentences and approximately 10K word forms. From the full-phrased morphological output clusters has been formed automatically viewing same root word in the morphological feature structure. The system reported an accuracy of 74.6%.

4. Conclusion

The present experiment has been successful and ended up with a fruitful result. We have tried this algorithm in Information Retrieval experiments. Any evaluation has not been done yet. We are now planning to implement some other hard clustering technique to increment the performance of the present system.

[2] http://ltrc.iiit.ac.in/ILMT/

References

1. Sajib Dasgupta and Vincent Ng. In the journal of Language Resources and Evaluation (LRE), 2007, published by Springer.

2. Karen Kukich . Techniques for automatically correcting words in text. In the ACM Computing Surveys, 1992, pages 377-439.

3. Sandipan Sarkar and Sivaji Bandyopadhyay. Study on Rule-Based Stemming Patterns and Issues in a Bengali Short Story-Based Corpus. In ICON 2009.