A Strategy for Morphological Analysis and Synthesis of Bangla

Abstract

With the advent of language technology, new areas of research such as NLP, MT, Question Answering System and Information Retrieval System have sprung up. In almost all the areas, an analysis of the language is seriously needed. Morphological Analyzer is a tool to analyze and identify every morpheme of a word. Bangla is a highly inflectional language with approximately 700 root verbs (single), and different forms for noun and pronoun. However, the root in both noun and verb take suffixes depending upon factors like syllable structure of the root, whether it is animate or inanimate etc. Since the primary aim is to build a quality machine translation system from English to Bangla, the morphological analysis of both the source and target languages becomes necessary. This paper deals with the morphological analysis of Bangla, and the problems related to it from the perspective of Machine Aided Translation System using AnglaBharti Technology.

Introduction

Morphology focuses on patterns of word formation within and across languages, and attempts to formulate rules that model the knowledge of the speakers of those languages. Thus morphological analysis is found to be concerned with the analysis and generation of word forms, and deals with the internal structure of words and how words can be formed. It may be mentioned that morphology plays an important role in the applications such as spell checking, electronic dictionary interfacing and information retrieving systems etc. In these applications, it is important that words, which are only morphological variants of each other, are identified and treated similarly. In natural language processing (NLP) and machine translation (MT) systems we need to identify words in texts in order to determine their syntactic and semantic properties.

General Discussion on Morphology

Morphological components are needed in parsing and/or generating natural language systems. Morphology can be defined as an internal structure of words. A Morphological Analyzer breaks a word into its root word and associated morphemes.

Morphological analysis can be broadly divided into two heads viz. Inflectional Morphology and Derivational Morphology.

Inflectional Morphology

In Inflectional morphology the basic meaning and part of speech of the root word and morphed word are the same. By adding prefix or suffix the root changes its grammatical features except part of speech like number, gender, tense and aspect of the verb and the like.
In inflectional morphology, an affix guli , tara etc. combines with a root to contribute a new member in the same grammatical class of the root.

Example 1: (Noun → Noun):
Chele 'boy'+ *guli* 'plural marker'→ *Cheleguli* 'the boys'

Example 2: (Adjective →Adjective):
kShudra 'minuscule'+ *tara* 'comparative suffix' → *kShudratara* 'comparatively minute'

Derivational Morphology

In derivational morphology however, after the addition of affixes to the root, the root generally changes its part of speech. Here, new words are derived from the root word through morphological changes.

In derivational morphology, an Affix bhAbe, bAna etc. combines derivationally with a root to contribute a new member in a different grammatical classes.

Example 3: (Adjective → Adverb)
bhAlo 'good'+ *bhAbe* 'manner'→ *bhAlobhAbe* 'in a good manner'

Example 4: (Noun→ Adjective)
rUpa 'beauty'+*bAna* 'adjectival suffix'→ *rUpabAna* 'handsome man'

However, recent study in this field has shown that there does not exist any watertight boundary between inflectional and derivational morphology.

Morphology can further be classified as linear and non-linear depending upon the type of structural changes of the root and other morphemes during the process of affixation. In linear morphology, affixes are directly added to the root without changing its internal structure. In non-linear morphology however, the internal structure of the morphemes changes during the process of affixation. Generally, the processes of pluralization and adjectivization fall under the category of linear morphology whereas Semitic languages and derivational morphology in Sanskrit features non-linearity. In Bangla language, the nominal and pronominal systems follow linear pattern whereas the verbal system features some non-linear characteristics.

Some Linguistic Aspects of Bangla

Although Sanskrit has been called the mother language of majority of Indian languages including Bangla, Bangla has also been highly influenced by the syntactic structure of Persian also. The Standard Colloquial Bengali (SCB) (which is the area of our study) and other dialects have originated from the so-called *sAdhu bhAsA* (the chaste language). Unlike Hindi, Bangla is not a gender-number language i.e. here the form of the verb does not get changed based on the gender and number of the subject or object. In Bangla, it does change with respect to tense, aspect, modality and person. As Bangla is an inflectionally rich language, we have 56 inflected form of a single root verb in Bangla

Maintaining the features of non-linear characteristics, sometimes a root verb changes its form when certain suffixes are added to it. For *example*, the verb *khA* 'to eat' when followed by the present continuous, Ist person suffix chchhi it becomes *khAchchhi*. On the other hand when followed by present perfect Ist person suffix *eYechhi*, it becomes *khe*Yechhi. So, by adding the suffix *eYechhi*, the root *khA* changes to khe and becomes *kheYechhi* , not *khAe*Yechhi which is clearly a non linear trend.

Bangla has got a number of classifiers that are added to the nouns. Among these, *tA ,ti, khAnA,*

khAni indicate singularity on one hand and on the other hand. *gulo, guli, sakala* etc indicate plurality. In this way, *chheletA* indicate 'a boy' and *chhelegulo* means 'the boys'. Featuring linguistic peculiarity, sometimes the singular suffix *tA* is also added with uncountable nouns like *jala* 'water' as in the sentence *jalatA AmAke* dAo 'give me the glass of water'. Here, *jalatA* indicates the container containing the water. This is a common trend of colloquial Bangla and it points out to the morphological richness of Bangla.

The suffix *era* is the marker of possessiveness or obliqueness. So *mAYera* means 'of mother'. The suffix *e* and *o* are used for specifying and hence they are generally referred to as specifier. These suffixes stand for 'only' and 'also' respectively. They take their place at the end of a noun or verb after addition of every possible suffix.

Noun Morphology

In Bangla nouns are inflected for case, including nominative, objective, genitive (possessive), and locative. The case marking pattern for each noun being inflected depends on the noun's degree of animacy and number.

Verb Morphology

Bangla Language contains a lot of verbs in which the core part is called root. In another way if we split the verbs we get two parts *Root* and *Suffix.* For example in the verb *kare* meaning 's/he does' has two parts: the root *kar* and the personal *suffix e*.

Morphological Analysis of Bangla verbs

Morphological analysis is applied to identify the actual meaning of the word by identifying suffix or morpheme of that word.

Bangla Roots

Every word is derived from a root word that may have the different transformations. This happens because different morphemes are added with it as suffixes. Therefore, the meaning of the word varies for its different transformations. For example, *medhA*(noun) 'intelligence' *medhAbI*(adjective) 'intelligent'.

Verbs divide into two classes: finite and nonfinite. Non-finite verbs have no inflection for tense or person, while finite verbs are fully inflected for person (first, second, third), tense (present, past, future), aspect (simple, perfect, progressive), and honor (intimate, familiar, and formal), but not for number. Conditional, imperative, and other special inflections for mood can replace the tense and aspect suffixes. The number of inflections on many verb roots can total more than 200.

Handling of different morphological elements in the MAT System

In the AnglaBharti System different tables have been used for handling verb, noun, preposition (postposition in Bangla), adjective, adverb and the like. Among these, verbs and nouns have been divided into some paradigms or categories according to their grammatical function.

Handling of verb morphology in AnglaBangla System

Correct translation output in any Machine Translation system is dependent on many factors. AnglaBangla system has the capability of generating sentences with different tense, moods, persons etc., each of which requires modification of the main verb of the input sentence. Such a process requires morphological

synthesis of verb. For this, verb roots have already been categorized and each category has been identified and implemented by means of paradigm numbers. Paradigm is basically a table for each category of verb providing all forms of that category. For example, the two verb roots *'kara'*(to do) and *'chala'*(to walk) have the same paradigm number as their inflected forms for different tenses and persons are the same. So in the verb paradigm table only one form has been kept instead of two. In this way paradigm tables have been generated for different part-of-speech. With the help of the verb paradigm table, generation of different forms of verb in simple past, present and future tense have been handled. In the Bangla verb morphological synthesizer, we have identified 34 paradigms for Bangla verbs; those can cover all the 730 number of Bangla verbs. Primarily total number of paradigm was 11, but after that, all the finer differences of the different verbal forms have been taken care of and correspondingly categorization has been increased from 11 to 34. At the time of creating the paradigm file honorific and non-honorific forms of Bangla verbs have been taken care of.

Like other languages, Bangla also have some irregular verbs, like *'yA'*(to go), *'Asa'*(to come) etc. For different persons and tenses inflected forms of these verbs are much different from the other root verbs. These verbs cannot be categorized with other verbs. In the Bangla text generator, the irregular verbs are handled in a different manner, not by paradigm tables. In AnglaBangla, one exceptional table has been maintained for these types of verbs, where all the forms of the particular root verbs are stored.

Again, for command or request type of sentences, the Bangla verbal forms are completely different from other. So, verb morphological synthesizer also considers these types. For example, in case of two English sentences, "Please do this job" and "**I do this job**", the first one is request type of sentence and the second one is affirmative type of sentence. But for both the cases same do verb has been used in English, but the corresponding Bangla forms are different for these two types of sentences, although the main verb in English is same. For the first sentence do will be translated as 'karuna'(honorific) or 'kara'(non-honorific) and for the second type it should be 'karachhi'. So, this needs implementation of special rules in the text generator.

Tense, aspect and modality of the verbs are handled in a separate file. Modal verbs like can, could, should, may etc are handled in this file. Aspects of verbs like continuous, perfect for all the tenses are taken care of here. The file contains a structure, which has five fields. These are i) Finiteness- this field gives information about the sentence type, like, normal, command, request, let etc. ii) Auxiliary verb iii) Main verb type iv) Phrase field and v) Suffix. Suffix field contains the suffixes which will be concatenated with the root verb.

As already mentioned, in AnglaBharti, root verbs and their paradigms have been dealt with in a file. In accordance with the AnglaBharti framework, root verbs of Bangla like khA(to eat), gA(to sing), kara(to do) etc have been assigned their paradigm numbers depending on their grammatical behaviour and also on the morphological changes of the root for taking the required inflection. In other words, for verbs like 'khA', the inflected form for past perfect is *'kheYechhilena'* .So, for adding the inflectional suffix *echhilena* with *khA,* the last character of

the root verb; i.e *'A'* is to be deleted. Only after the deletion of the last character, can the suffix *'echhilena'* be added to the root. Thus, it becomes important to keep in mind the number of characters to be deleted for adding a particular suffix to a particular Bangla root verb. Verbs in AnglaBharti have been divided into 11 groups or paradigms with provision for increasing the number of paradigms according to the verbal structure of specific languages. In case of Bangla, initially attempt was made to map all the Bangla verbs within these 11 paradigms. But it was soon found that the prescribed 11 paradigms were not sufficient for taking care of all Bangla verbs. For example, the paradigm no.10 was used for Bangla root verb *'mApa', 'pAra'* etc. It was observed that the indefinite and continuous forms for both verbs were same, as inflections were *'chhi'* for 1st person present continuous or *'chhilAma'* for 1st person past continuous etc. But problems were occurred in the case of present perfect and past perfect where inflections were added after deletion of some characters from root verb. Before increasing the number of verb paradigm, it was found that for 1st person present perfect, the inflectional suffix *'echhi'* was added after deletion of one character from root verb in paradigm no. 10. So the verb forms for both were occurred as *'mApechhi'* and *'pArechhi'* which were wrong. The right verb forms for above mentioned roots are *'mepechhi'* and *'perechhi'*. For applying the right form, *'epechhi'* is used after deletion of three characters, i.e *'Apa'* from root verb *'mApa'* and *'erechhi'* is used after deletion of three characters i.e *'Ara'* from root verb *'pAra'*. Thus it is impossible to map the both verbs in same paradigm no. That's why new paradigm no. 11 is applicable for verb root *'mApa'* and paradigm no. 15 is used for verb root *'pAra'*. Same formula was applicable for root verb *'kATa'* (paradigm no.13), *'thAka'* (paradigm no.23), *'thAma'* (paradigm no.24), *'nAcha'* (paradigm no.26) etc. So, after a careful analysis of Bangla verbs, the number of verbal paradigms has been increased from 11 to 34.

Conclusion

Morphology is an important element of study in the development of a machine translation system. The degree of performance of a MT system depends a lot on a good morphological analyzer and synthesizer. This paper attempts to provide some ideas about the morphological synthesis involved in the AnglaBangla Machine Translation System. This is only a stepping-stone in the area of machine translation and it can be hoped that it would facilitate the development of MT system in other languages also.