Chapter 1

Introduction

(This chapter contains problem definition, objective of research, challenges for building Morphological Analyser and organization of the thesis)
Linguistically, morphology means analysis of word formation and their relationship to different words in the language. Morphology of a word describes the structure of words and parts of words, for example, stems, root words, prefixes, and additions. Morphology additionally informs about parts of speech, intonation and stress, and the ways context can change a word's pronunciation and meaning.

Morphological Analysis is a process by which we can find out the base form of an inflected word. Morphological Analyzer is computer program for analyzing inflected word to find the based form along with its additional information’s (affix information)

Morphological analyzer is one of the required tools for processing text analysis of any human language automatically. The morphological analyzer returns root word (base form) and its peripheral information.

Example in Assamese Language

- ოჰ (verb) + ᴮ (past tense marker)
- ṭ (noun) (number) + Ṇ (definite case marker)
- ო (noun) (Feminine Gender) + ᶇ (indefinite plural marker)
- ῴ (verb) (3rd person)+ ᶇ (nominative case marker)
1.1. Problem Definition & Objective of Research

Numerous indigenous languages of India especially the north – east Indian languages have received less awareness from computational linguists than the other standard Indian and European languages. Due to less numbers of available standard digital information of Assamese Language, it is very complex to build a complete morphological analyser including all domains against other European Languages. Word formation process in Assamese Language is very complex due to agglutinative nature, for which it is very difficult and challenging task for building morphological analyser using a computational approach.

For the successful implementation of any morphological tool, an essential requirement is standard and comprehensive linguistic description of the language along with a good numbers of standard digitized corpuses for that language.

1.1.1. Motivation to build Morphological Analyser

Morphological analyser is one of the imperative tools for processing of text analysis of any human language automatically. It examinations the word shapes in a sentence and distinguishes the root word and its components.

Morphological analysers can be used for many applications. Followings are some applications where morphological analyser is the primary requirement to continue to the applications [38]

- Search engines for retrieving the documents from a keyword.
The absence of morphological analyser tool hampers the development or the improvement above-mentioned applications.

The main objective of my research is to develop a successful morphological analyser for Assamese Language. For which I am taking the help of Statistical Model & Finite State Techniques.

Followings are divisions of my research objectives

- **Linguistic knowledge acquisition**: Acquiring the morphological knowledge of Assamese Language in computational perspective for developing morphological application

- **Efficiency & Sufficiency**: Morphological analyser is already developed for many languages using Finite State Techniques. We are also adopting FST based technique and for analysing few features we are using statistical technique. So it is sufficient to handle different irregularly found in Assamese Language and can build an efficient Morphological Analyser
– **Coverage**: An FST based technique can cover a larger no of word formation of a language. A Statistical model can give good performance best on provided knowledge to the system. So, sufficiently a large range of Assamese Language word forms can be analysed using these techniques.

– **Effort**: We are giving all effort to implement a successful Analyzer for Assamese language using our hybrid analyzer (combining finite state technique and statistical method).

### 1.2. Challenges for developing Morphological Analyser

Major Challenges for developing Morphological Analyser for Indian Language and Highly Agglutinative Language are

**From Language aspects**

– **Morphological variation**: Most of the Indian languages are using compound formation by adding more than one root word to create a new word. Sometimes partially combining more than one root word to create new word. Tense, number, mood markers, gender etc. information’s are added with a word by adding affixes.

– **Use of foreign word in the language**: A language may contain loan words. For Example: school/স্কুল, telephone/টেলিফোন, mobile/মোবাইল etc.

– **Presence of polysemy words in a language**: Presence of polysemy words in a language creates problems for defining meaning and parts of speech tag for a word. For example, Assamese word “বল” can be used as
Reduplication of words: Reduplicative nature of a word may present in a language for example in Assamese Language: লাহেলাহে, চোলাঙোলা, পানীছোলী, ভালা ভালা etc.

Changing characters or hiding characters: In Assamese Language, after affixation or compound formation some characters are hidden in the final form or some characters are changed to other characters.

For example: নাটে – নাঁ, (এ, কৃষ্ণার্থন –কৃষ্ণ) অর্জুন)

Changing of the parts of speech tags: Based on the on the position of the word present in a sentence; the word POS category may be changed. “তুমি এতিরা সাতোঁবা” Here the word “সাতোঁব” is placed as verb in the sentence. But in “সাতোঁব একিচ ভাল বায়াম” The word “সাতোঁব” is placed as noun.

Form Computational aspects

- Based on the problem selecting supervised or unsupervised model
- Collecting sufficient data for supervised model.
- Marking different tags and parameters for training text corpus.
- Analysing Morphotactics of a language model and rewriting the rules.
- Developing an identical model or finding a matching algorithm/model, which will be suited for all the domains of a language. Since it is unrealistic to build
a 100% efficient model which will cover all the domains all the aspects of a language.

1.3. Organization of the Thesis

The organization of the thesis is as given below:

Chapter 1: This chapter contains problem definition, objective of research, challenges for building Morphological Analyser and organization of the thesis

Chapter 2: Describes the overview of Natural Language Processing, Morphology, linguistic history and linguistic characteristics of Assam.

Chapter 3: Shows different literatures related to the morphology, morphology of Indian Languages including Assamese Language.

Chapter 4: Describes the detail morphological analysis of English language context, Indian Language Context and Assamese Language Context.

Chapter 5: Describes required resources for developing a morphological Analyser and the current design consideration

Chapter 6: Describes Assamese Morphological Analyser’s Architecture & Prototype and implementation details for finite-state transducer.

Chapter 7: Describes the implementation details of hybrid architecture using finite-state-transducer and statistical model.

Chapter 8: Shows the results and analysis of the research.

Chapter 9: Describes the concluding remarks and the future work of the research.