Current Studies on Morphological Analyzer: In context of Assamese Language

Mirzanur Rahman[1], Shikhar Kumar Sarma[2]
Department of Information Technology[1, 2], Gauhati University[1, 2]
Email: mirzanurrahman@gmail.com[1] ,sks001@gmail.com[2]

Abstract- Morphological Analysis is a branch of linguistics in the domain of Natural Language Processing. Morphology studies the word structure and formation of word of a language. In current scenario of NLP research, morphological analysis techniques have become a popular for all the languages. Because, for processing any language, morphology of the word should be first analyzed. In India, most of the languages contain very complex morphological structure. The aim of this paper is to give a summarize view of available literature information so that the researchers can look into these techniques and try to develop better ones.

Index Terms- Assamese Language, Morphology, Natural Language Processing, Review, Assamese MA

1. INTRODUCTION

A natural language is known as a language that is spoken and written for general-purpose communication. For making a natural language understandable to a computer, we need several information including word grammar and sentence structure of that language. The processing of this information's by a computer is known as natural language processing (NLP). So the area of NLP research covers the how natural human languages can be automatically generated and understand by a computer.

Morphological Analysis is one part of NLP research which studies the structure of words and word formation of a language.

Words in a language can be divided into many small units. In morphology the smaller meaningful units are called morphemes. The problem of recognizing different morpheme in a word is known as morphological parsing or Morphological Analysis [18]. For example in English language

- **Boys=Boy +s**
  - Root : Boy (category- noun)
  - 's' (indefinite plural marker)

In the above example, the word "Boys" is a combination of two morpheme "Boy" and "s" (root word "Boy" and suffix "s"). If a morphological analyzer analyses a word, it should give both root word and affix added with it and some other information's like tens, case marker, gender, number, person and other relevant morphological information etc as output.

In the information retrieval domain, mapping from Boys to boy is called stemming.

There are two broad classes of morphemes: stems and affixes. The stem is the 'main' morpheme of the word, supplying the main meaning, while the affixes add 'additional' meanings of various kinds.

Affixes are further divided into

- **Prefixes:** These are placed in front of the stem
  Example: Unbuckle =Un+ buckle
- **Suffixes:** These are added after the stem
  Example: Boy= Boy + s
- **Infixes:** These are inserted in the middle of the word.
  English language has no true infixes, but the plural suffix -s behaves something like an infix in unusual plurals.
  Example: Mothers-in-law= Mother + s + in-law
- **Circumfixes:** These are placed in both side of the stem, i.e. front and end.
  Example: Unbelievable=Un + believe +able

Prefixes and suffixes are often called concatenative morphology since a word is composed of a number of morphemes concatenated together.

Nonconcatenative morphology, also called discontinuous morphology is a form of word formation in which the root is modified and which does not involve stringing morphemes together [17].

In English while plurals are formed, it use the suffix –s normally. But there are certain words which use nonconcatenative processes for their plural forms. For example:

Singular form		**Plural form**
bite	→	bit
sing	→	sang
give	→	gave
feel	→	felt

2. PRIOR ARTS

In this section we describe the important and relevant research works done in the field of morphological analysis.

Morphological Analysis: In terms of Indian Language

In Indian language context, a number of rule based stemmers have been reported. Among these, we try to summarize some of the reported work.

- Hindi Language:

  [1] used a hand crafted suffix list and stripped off the longest suffix for Hindi and reported 88% accuracy for their approach using a dictionary of size 35,997.
  The work of [3] focused on some heuristic rules for Hindi and reported 89% accuracy.

- Bengali Language:

  The work reported by [2] learned suffix stripping rules from a corpus and used clustering to discover the nearest class of the root word for Bengali, English and French. They described a centroid based approach that rewards the longest common prefix to form similar word clusters based on a threshold value.
  An open-source morphological analyzer for Bengali Language using finite state technology was developed by [8]. This is the first open source attempt in creating a fully-functional morphological analyzer and the system is currently in under development stage.

- Punjabi language:

  Reported work of [4] for Punjabi used a dictionary of size 52,000 and found 81.27% accuracy using a brute-force approach.

- Marathi Language:

  [5] described a hybrid method (rule based + suffix stripping + statistical) for Marathi and found 82.50% precision for their system.

- Tamil & Malayalam Language:

  Work in Malayalam [6] used a dictionary of size 3,000 and reported 90.5% accurate system using finite state machines.
  [7] developed a rule based Morphological Analyzer and generator for Tamil using finite state transducer called AMAG. The performance of the system is based on lexicon and orthographic rules from a two level morphological system. The system consists of list of 50000 nouns, around 3000 verbs and a relatively smaller list of adjectives.
  Based on suffix stripping and suffix joining approach, using a bilingual dictionary, a Malayalam morphological analyzer and a Tamil morphological generator have been developed by [9] the developed analyzer and generator were used for Malayalam - Tamil machine translation.

- Kannada Language:

  [10] Developed a rule based MAG for Kannada language using finite state transducer (FST). The proposed MAG is capable of analyzing and generating a list of twenty thousand nouns, around three thousand verbs and a relatively smaller list of adjectives. The uniqueness of the proposed MAG is its capacity to generate and analyze transitive, causative and tense forms apart from the passive constructions, auxiliaries and verbal nouns.
  Using rule based with paradigm approach used in [10] , [11] proposed a morphological analyzer and generator for Kannada language . They used Trie as a data structure for the storage of suffixes and root words. The disadvantage of Trie is that it consumes more memory. As a result it can handle up to maximum 3700 root words and around 88K inflected words.

Morphological Analysis: In terms of Assamese Language

For Assamese language also we have found some of the reported work for morphological analysis. In this section we will try to summarize all the reported work related to Assamese morphological analysis.

- In [12], the authors have presented building Morphological Analyzers using the Suffix Stripping method for the four languages – Assamese, Bengali, Bodo and Oriya. In the proposed mechanism they have deals with only inflectional suffixes. The method involves identifying individual suffixes from a series of suffixes attached to a stem/root, using morpheme sequencing rules.

  In the approach the analyzer analyses the inflected form of a word into suffixes and stems by using a root/stem dictionary (for identifying legitimate roots/stems), a list of suffixes, comprising of all possible suffixes that various categories can take (in order to identify a valid suffix), and the morpheme sequencing rules.

  The designed analyzer gives three types of outputs:

  - The Correct analysis: if complete match of suffixes, rules and the existence of the analyzed stem/root in the root dictionary, then this may be obtained.
  - Probable analysis: This is obtained on the basis of either a matching of the suffixes and rules, even if the root/stem is not found in the dictionary or a matching of the suffixes, but not any supporting rule or existing root in the dictionary.
  - Unprocessed words: These are the words which have remained unanalyzed due to either absence

of the suffix in the suffix list or due to the absence of the rule in the list.

The strong point of the proposed approach is that, it is highly efficient in case of agglutinative languages.

The weak point of these methods is that it directly related to root word dictionary size. The authors get 50 % coverage for 7000 to 8000 root entries.

- In [13], the authors have presented A Suffix-based Noun and Verb Classifier for an Inflectional Language. In the proposed mechanism they have consider only the morpho-syntactic properties of Assamese words. Assamese words can be categorized into inflected classes (noun, pronoun, adjective and verb) and un-inflected classes (adverb and particle.

  In their method, they used EMILLE Assamese text corpus (5300 sentences), jointly developed by Lancaster University and CIIL-Mysore. They tokenized the corpus by considering white space as word separator and punctuations (.,?,!) as sentence terminator.

  The proposed method follows three basic steps to tagged tokenized text.

  - Brute-force determination of suffix sequences: There they obtain all possible sequences of noun suffixes.
  - Suffix sequence pruning: In this step they filter out the non-valid suffix sequences. Though a number of suffix sequences can possibly occur after a root word.
  - Suffix stripping: In this step, they identify the noun and verb roots based on the single suffix that occurs immediately after the root.

  Though the main aim of the above methods is to build POS tagger for Assamese language, they have implement morphological analyzer using Suffix stripping methods for their work

- In [14], the authors describe an approach to unsupervised learning of morphology from an unannotated corpus for Assamese Language in their paper “Acquisition of Morphology of an Indic Language from Text Corpus”. In their paper they have present & elaborately discussed an unsupervised method for acquisition of Assamese morphology from a text corpus.

  For experiment, they have used text chunks from different sources. They used lexicon and the set of suffixes and suffix sequences obtained from corpus which was collected from 525 newspaper articles that include general news, sports news, and editorial articles. For testing, they ran their process over 84 other newspaper articles totalling 32,271words from the same newspaper source, and 66 articles from the Emille corpus for Assamese (http://www.ling.lancs.ac.uk/fass/projects/corpus/emille/)

  According to their observation, this method provides an analysis for over 95% of the words for all newspaper articles, that is, either the words are decomposed or conclusively declared as root words and small fraction of words the method does not provide any analysis. In case of the Emille corpus, for most articles this ratio is over 92%.

  The strong point of the proposed unsupervised approach is that, this is the initial work towards unsupervised morphological analysis and it is very suitable for Assamese language. This approach, acquire the suffixation morphology of the language from a text corpus of about 300,000 words and build a morphological lexicon. The F-measure of the suffix acquisition is about 69%.

- In [15], the authors have presented suffix stripping approach, where they add a rule engine which generates all the possible suffix sequences for analyzing morphology of a word. They got 82% accuracy with a root-word list of size 20,000 approximately with this method.

  In their work they have presented two methods for analyzing morphology of the word.

  - In Method 1, they first read a line from the corpus file. Then extract words (token)from the line, clean the token, that is remove punctuation marker attached with token if there is one.

    In next step Look up suffix-list generated manually from the end of the token. If matched with the suffix-list extract and exit from the process. This process continues up to the last word of the corpus

    The authors claimed that they found, 61% words were correctly stemmed.
  - In Method 2, they have used a root word dictionary after tokenizing the word. If a dictionary entry matches with the token, mark token as root word and exit otherwise execute the next step as method 1.

    Using this method, the authors found that, 82% words were correctly stemmed.

    The strong point of this method is that, here authors introduced a root word dictionary with suffix stripping method, which increases accuracy of word stemming

- In [16], the author combines a rule based algorithm and HMM based algorithm. Where rule based algorithm is used for predicting multiple letter suffixes and an HMM based algorithm for predicting the single letter suffixes .This added method can predict morphologically inflected words with 92% accuracy.

By this method, the authors get 91% accuracy for inflected word with single character suffixes. The strong point of this method is that, it combines both rule based algorithm for predicting multiple letter suffixes and an HMM based algorithm for predicting single letter suffixes which leads to a higher overall accuracy of 92% compared to 81% for previously published methods for Assamese.

- In [19] Utpal Sarma proposed an unsupervised method for learning morphology of a word in his Ph.D thesis "Unsupervised Learning of Morphology of a Highly Inflectional Language"

3. CONCLUSION

In this paper work, we have presented a survey on different developments of morphological analyzer for some of the Indian languages. We have found that almost all approaches used rule based method for morphological analyzer and generator. For Assamese Language also, the literature reveals that most of the works are based on supervised suffix striping method. Only one or two [14, 19] reported method has applied unsupervised technique for analyzing morphology of the word.

REFERENCES

[1] A. Ramanathan and D. D. Rao. "A lightweight stemmer for Hindi". In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL), on Computatinal Linguistics for South Asian Languages, 2003.

[2] P. Majumder, M. Mitra, S. Parui, G. Kole, P. Mitra, and K. Datta. "YASS: Yet another suffix stripper. ACM Transanctions and Information Systems", 25, 2007.

[3] A. Pandey and T. Siddiqui. "An unsupervised Hindi stemmer with heuristic improvements. In In Proceedings of the second workshop on Analytics for noisy unstructured text data", page 99ˆa ˘A¸S105, 2008.

[4] D. Kumar and P. Rana. "Design and development of a stemmer for Punjab"i. International Journal of Computer Applications, 11, 2011.

[5] M. M. Majgaonker and T. J. Siddiqui. Discovering suffixes: "A case study for Marathi language. International Journal on Computer Science and Engineering", 04:2716–2720, 2010.

[6] V. S. Ram and S. L. Devi. Malayalam stemmer. "In Morphological Analysers and Generators", pages 105–113, 2010.

[7] Dr. A.G. Menon, S. Saravanan, R. Loganathan and Dr. K. Soman,, "Amrita Morph Analyzer and Generator for Tamil: A Rule Based Approach", Tamil International Conference, 2010, Coimbatore, India.

[8] Abu Zaher Md. Faridee and Francis M. Tyers, "Development of a morphological analyser for Bengali", Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation, p. 43-50, Alacant, Spain, November 2009.

[9] Jisha P.Jayan, Rajeev R R and Dr. S Rajendran, "Morphological Analyser and Morphological Generator for Malayalam - Tamil Machine Translation", International Journal of Computer Applications (0975 – 8887, Volume 13– No.8, January 2011.

[10] Ramasamy Veerappan, Antony P J, Saravanan S and Soman K P: "Rule based POS tagger for Kannada language using Finite State Transducer ", International journal on Computer Application (IJCA), ISBN: 978-93-80746-92-0, 2011.

[11] Shambhavi. B. R, Dr. Ramakanth Kumar P, Srividya K, Jyothi B J, Spoorti Kundargi, and Varsha Shastri G, " Kannada Morphological Analyser and Generator Using Trie", International Journal of Computer Science and Network Security (IJCSNS), VOL.11 No.1, January 2011

[12] Mona Parakh and Rajesha N, "Developing Morphological Analyzer for Four Indian Languages Using A Rule Based Affix Stripping Approach", Linguistic Data Consortium for Indian Languages, CIIL, Mysore, 2011.

[13] Navanath Saharia, Utpal Sharma and Jugal Kalita, "A Suffix based Noun and Verb Classifier for an Inflectional Language" International Conference on Asian Language Proceesing(IALP-10), China, 2010

[14] Sharma, Utpal and Kalita, Jugal K and Das, Rajib K. "Acquisition of Morphology of an Indic Language from Text Corpus". ACM Transactions of Asian Language Information Processing (TALIP), vol 7, no. 3, article 9, p 1-33, August 2008.

[15] Navanath Saharia, Utpal Sharma and Jugal Kalita, "Analysis and Evaluation of Stemming Algorithms: A case study with Assamese" Proceedings of the International Conference on Advances in Computing, Communications and Informatics, Pages 842-846, Chennai, 2012

[16] Navanath Saharia, Kishori M. Konwar, Utpal Sharma and Jugal Kalita, "An Improved Stemming Approach Using HMM for a Highly Inflectional Language", Proceedings of 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), Pages 164-173, Samos, Greece, 2013

[17] (2014) Nonconcatenative morphology, Available: http://en.wikipedia.org/wiki/Nonconcatenative_morphology

[18] James Allen ,Natural Language Understanding, Second Edition, Pearson Education India, ISBN: 8131708950

[19] Utpal Sharma, Unsupervised Learning of Morphology of a Highly Inflectional Language, Phd. Thesis, 2006