Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 12, December -2017

A Survey on Morph Analyzers for Indian Languages

Saroj Kumar Jha[1], Amit Kumar Jha[2], Piyush Pratap Singh[3]

Dept. of Computational Linguistics, School of Languages,
Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalay, Wardha (MS)

Abstract: *Computational Morphology of natural language is a challenging task incorporated from Computational Linguistics (CL), an applied area of NLP diverged from Artificial Intelligent (AI). Heading towards the assembling a robust Machine Translation (MT) system, Morphological analyzer (MA) is the vital stages where the words are picked to analyzed up to the minimal meaningful unit of that language. The ample of morphological analyzer applications have been developed which covers all the possible approaches of machine learning which clearly stats the output of the application, procedure of development and possible up-gradation throughout the time. This paper is a survey incorporated with all the survey of research work done on MA and having an aim to discover a most feasible approach could be elected for the development of Maithili Morphological Analyzer. Keeping such vision, all the research papers of MA-IL is reviewed and diagnosed at various levels, the most used mechanism would be applied with possible modification to meet our aim and also analyze it with all the related morpheme features. This survey is absolutely motivated based on all the MA-IL applications having various approaches like Algorithm, accuracy, and attached features. And up to what extent the elected approach would help to develop Maithili MA with robust accuracy.*

Keywords: *CL, NLP, MT, MA, IL.*

Introduction

Morphological analysis is one of the significant processes of overall development of Machine translation for Indian Languages. Developing the MT for Maithili to any other IL without morphological analysis is an insignificant effort. The morphological analysis is mostly covered in two major parts one is inflection and the other is derivation. Derivational Morphology is also known as Word formation process which contains the formation process attached with Affixes (Prefix & Suffix). This paper is a quick review about all the morphological analyzer of Indian languages till date. It contains the small process of covering all the MA developed for Indian language and what sort of approaches has been used.

State of Arts

[1] The research paper on *"Unsupervised Improvement of Morphological Analyzer for Inflectionally Rich Languages"* written by (Bharti, Akshar. et. al. 2001). This paper presented an algorithm for unsupervised learning of morphological analysis and generation of inflectionally rich language like Hindi, given a low coverage morph and a corpus of raw text. The result of the algorithm are encouraging with the coverage of primitive morph going up from 32% to about 63% and that of an advanced morph going up from 96% to about 97%.

[2] The paper on *"Developing a Finite State Morphological Analyzer for Urdu and Hindi"* is written by (Bogal, Tina. et. al. 2007). In this paper we have introduced and addressed a number of issues that arise in the process of building a finite-state morphological analyzer for Urdu. Our approach allows for an underlying similar treatment of both Urdu and Hindi via a cascade of finite state transducers that transliterates the very different scripts into a common ASCII transcription system. We further explored reduplication in Urdu, again basing ourselves on solutions proposed with respect to XFST and show how differing reduplication patterns in Urdu/Hindi can be dealt with elegantly with the finite-state methods proposed by B&K. Finally, we addressed some potential ambiguity problems and discussed different ways of solving them.

[3] The research paper on *"Hindi morphological Analysis & inflectional generator for English to Hindi translation"* written by (Singh, Pawan Deep et.al. 2008). This paper primarily discussed on analysis of nominal inflectional in Hindi within the framework of distributed morphology. The authors discus about the analysis categories, inflectional classes, morphological processes operating at syntax, the distributed vocabulary items & readjustment rules for Hindi nouns and analysis the experimental results we obtained from the system developed and as well as the accuracy of the same. In this system words are tokenizing using NLTK tool, then a Stanford parser are used to tagging of words. The system was able to completely analyze in most cases accurately. The system failures were driven primarily by external factors.

[4] The research paper on *"Hindi Morphology Analyzer & Generator"* written by Goyal, Vishal et. al. 2008). This paper presents the morphological analysis and generator tools for Hindi language using paradigm approach for windows

platform having GUI. This project has been developed as part of the development of machine translation system from Hindi to Panjabi language. Hindi is very rich in inflectional morphology can be witnessed from the fact that is English usually there are maximum 7-8 inflected word forms of noun but in Hindi it can be up to 40 and even more than that. This morphological analyzer gives preference to the time taken to search for a word in the database to know its grammatical information & also accuracy of returned results. In the database used by this tool all the possible word forms of all root words are stored. Though it takes a bit more space but the search time is very less.

[5] The paper on *"Developing Morphological Analyzer for South Asian Languages: Experimenting with Hindi and Gujarati Languages"* is written by (Ashwani, Niraj. et. al. 2010). This paper is described on morphological analyzer for the Hindi and Gujarati language. In order to demonstrate our approach's portability to other similar languages, we present our experiments for Gujarati language. The paper presents a rule-based morphological analyzer where the rules are acquired semi-automatically from corpora. The experiment proposes an approach that takes both prefixes as well suffixes into account. Given an inflected Hindi word, our system returns its root form. It uses a dictionary, and a monolingual corpus to obtain suffix-replacement rules. The improvement is partially depending on the GRFL list which causes variation in result.

[6] The research paper on *"Hindi Derivational Morphological Analyzer"* is written by (Kanuparthi, Nikhil et.al. 2012). In this paper the authors present their Hindi derivational morphological analyzer. Their algorithm upgrades an existing inflectional analyzer to a derivational analyzer & primarily achieves two goals. First it successfully incorporates derivational analysis in the inflectional analyzer. Second, it also increases the coverage of the inflectional analysis of the existing inflectional analyzer. The authors pursued the five steps approach for building their derivational analyzer – studying Hindi derivations, derivational rules, finding majority properties, using Wikipedia data for confirming genuineness, Develop an algorithm for derivational analysis. The algorithm uses the principle of Porter's stemmer & Krovetz stemmer.

[7] The research paper on *"Development of Morphological Analyzer for Hindi"* is written by (Rastogi, Mayuri et. al. 2012). This paper is primarily concerned with the design of morphological analyzer for Hindi language. The input to this analyzer will be a Hindi word or sentence and after doing the proper analysis it will return the root word along with its feature as output. The features will have categories like part of speech, gender, number and person. Two approaches will be followed by analyzer- rule based and corpus based. The approach discussed in this paper has capability of working on either type of morphology.

[8] The research paper on *"Statistical Morphological Analyzer for Hindi"* written by (Malladi & Mannen, et. al. 2013). This work have stated that the comprehensive evaluation of PBA using the data from Hindi Treebank (HTB) and trained on it. After training on HTB the result is much higher from the existing available Morph Analyzer of Oracle and others. The accuracy after using HTB on Gender, Number, Person and Case (GNPC) is 84.16% for Hindi. The developed analyzer has acchived the accuracy of 82.03% for lemma, Gender, number, person, case, vibhakti and TAM.

[9] The research paper on *"Morphological Analysis for a given text in Marathi Language"* written by (Aditi Muley et. al. 2013). Morphological analyzer analyzes the given words and generates from the stem and its features (like affixes). This paper presents the morphological analysis for Marathi language using Rule-Based Approach.

[10] The research paper on *"Context Based Statistical Morph Analyzer and its Effect on Hindi Dependency Parsing"* written by Malladi, et al (2013). This paper has concluded that SMA is a robust state of art statistical morphological analyzer which out forms pervious analyzer for Hindi by considerable margin. SMA achieved an accuracy of 63.06% for *Lemma* (L), *Gender*, *Number*, *Person* and *Case* where as PBA and Morfette are 34.89% and 51.52% accurate respectively. With the predicted morphological attributes by SMA, we archived a label attachment score of 89.41% while without these morphological attributes the parsing accuracy drops to 87.75%.

[11] The paper on *"Morphological Analyzer for Hindi- A Rule Based Implementation"* is written by (Agarwal, Ankita. et. al. 2013) have discussed in this paper a Hindi morphological analyzer which is basically based on rule based approach but also utilize the corpus when exception occurs. The paper has incorporated all the possible rules for the different word formation of Hindi as described in the section 3 be it inflectional or derivational. The analyzer is performing better and the accuracy of the system is very high and the possible expectations are covered. The analyzer has been integrated with the word sense disambiguation so that the words having multiple senses could also be analyzed accurately.

[12] The paper on *"Morphological Analyzer for Malayalam: Probabilistic Vs Rule Based Method"* is written by (Rinju, O.R. et. al. 2013). This paper presents a Morphological Analyzer for Malayalam, by considering the noun and verb categories of a word. The proposed morph analyzer returns the morpheme along with the grammatical information such as Gender, Number and case information of noun and tense aspect for verb. A probabilistic and rule based

method is used for analysis using inflection and suffix list, which is created using look up tables. The result shows that the rule based method is more accurate than the others.

[13] The research paper on *"Statistical analyzer (SMA++) for Indian languages"* is written by (Srirampur S. et. al. 2014). This paper is the improved version of (SMA) described in Malladi and Mannem (2013). SMA++ predicts the gender number, person, case (GNPC) and lemma (L) of a given token. The SMA is further modified with some machine learning features in addition to some features sets based on the characteristics of Indian languages. This mechanism is applied basically on the four Indian languages viz. Hindi, Urdu, Telugu and Tamil and the accuracy was 85.87%, 79.16%, 86.81% and 78.97% respectively.

Description of All the Morph Analyzer

S. No.	Detail Description	Approach	Output (Apx.)
1.	Unsupervised Improvement of Morphological Analyzer for Inflectionally Rich Languages	Statistical Approach	96-97%
2.	Developing a Finite State Morphological Analyzer for Urdu and Hindi	XFST Approach	--
3.	Hindi morphological Analysis & inflectional generator for English to Hindi translation	Hybrid with NLTK	--
4.	Hindi Morphology Analyzer & Generator	Paradigm Approach	--
5.	Developing Morphological Analyzer for South Asian Languages: Experimenting with Hindi and Gujarati Languages	Rule Based	--
6.	Hindi Derivational Morphological Analyzer	Corpus Based Approach	--
7.	Development of Morphological Analyzer for Hindi	Statistical Approach	H-85.87%
8.	Statistical Morphological Analyzer for Hindi	Corpus & Rule Based	82.03%
9.	Morphological Analysis for a given text in Marathi Language	Corpus Based Approach	--
10.	Context Based Statistical Morph Analyzer and its Effect on Hindi Dependency Parsing	Statistical Approach	87.75%
11.	Morphological Analyzer for Hindi- A Rule Based Implementation	Rule Based	--
12.	Morphological Analyzer for Malayalam: Probabilistic Vs Rule Based Method	Probabilistic Vs Rule Based Method	82%
13.	Statistical analyzer (SMA++) for Indian languages	Statistical Approach	78.97%

Feasible Approach for Maithili MA

As the above analyzers are comparatively evaluated and found most of them are corpus based and rest are statistical. Even though we look for the ratio of functional analyzer at present, very few are working with high accuracy and rest of them functional according to its requirement. But if we talk of approaches used in most of the functional analyzers, the findings took us to corpus based approach. Although accuracy among the all the above analyzers statistically is far ahead from the others. Therefore, overall discussion lead us to use statistical approach but if we take the advancement in our analyzer, the combination of statistical and rule-based will provide a better output then pure statistical.

Conclusion

The throughout discussion about the survey to assemble Morphological Analyzer for Maithili lead us to use a Hybrid approach. Where it could be the combination of corpus based approach and machine learning technique. We will surly apply this and will watch over all the possible challenging aspects.

References

[1] Bharati Akshar, Sangal Rajeev, Bendre S.M., Kumar Pawan, "Unsupervised Improvement of Morphological Analyzer for Inflectionally Rich Languages"

[2] Tina B¨ogel, Miriam Butt, Annette Hautli, & Sebastian Sulger. Developing a finite-state morphological analyzer for Urdu and Hindi. Finite State Methods and Natural Language Processing, page 86, 2007.

[3] Singh Pawan Deep, Kore Archna, Sugandhi Rekha, Arya Gaurav, Jadhav Sneha. Hindi morphological Analysis & inflectional generator for English to Hindi translation, IJEIT, 2008.

[4] Vishal Goyal, Gurpreet Singh Lehal, "Hindi Morphological Analyzer and Generator", First International Conference on Emerging Trends in Engineering and Technology, USA, pp.1156– 1159, 2008.

[5] Niraj Aswani, Robert Gaizauskas, "Developing Morphological Analyzers for South Asian Languages: Experimenting with the Hindi and Gujrati Languages", Proceedings of the Seventh International Conference on Language Resources and Evaluation, Valleta, Malta pp.811-815, May, 2010.

[6] Nikhil Kanuparthi, AbhilashInumella, DiptiMisra Sharma, "Hindi Derivational Morphological Analyzer", Proceedings of the twelfth meeting of the Special Interest Group on Computational Morphology & Phonology, Canada, pp.10-16, June, 2012.

[7] Rastogi Mayuri, Khanna Pooja. Development of Morphological Analyzer for Hindi. IJCA, 2014.

[8] Malladi Deepak Kumar, Mannem Prashanth,. Statistical morphological analyzer for hindi. In Proceedings of 6th International Joint Conference on Natural Language Processing, 2013.

[9] Muley Aditi, et al., "Morphological Analysis for a given text In Marathi language", International Journal of Computer Science & Communication Network, Vol-4 (1), 13-17, 2014.

[10] Deepak Kumar Malladi and Prashanth Mannem. Context based statistical morphological analyzer and its effect on hindi dependency parsing. In Fourth Workshop on Statistical Parsing of Morphologically Rich Languages, volume 12, page 119, 2013.

[11] Agarwal Ankita, Pramila, Singh Shashi Pal, Kumar Ajai, Darbari Hemant,. *Morphological Analyser for Hindi – A Rule Based Implementation. In proceeding of International Journal of Advanced Computer Research* (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-4 Number-1 Issue-14 March-2014.

[12] Rinju, O., Rajeev, R., and Sherly, E., Morphological analyzer for malayalam: Probabilistic method vs rule based method, 2013.

[13] Srirampur Saikrishna, Chandibhamar Ravi, Mamidi Radhika., Statistical Morphological Analyzer for Hindi, 2014.

