A STRUCTURED ANALYSIS ON MORPHEME SEGMENTATION FOR AGGLUTINATIVE LANGUAGES

Dr. Ananthi Sheshasaayee[1], Angela Deepa.V.R[2]

[1]*Research Supervisor, Department of Computer Science & Applications, QMGCW, Chennai (India)*

[2]*Research Scholar, Department of Computer Science & Applications, QMGCW, Chennai (India)*

ABSTRACT

The learning of morphology through unsupervised methods helps to identify automatically affixes, morphological segmentation of words with engendered paradigms which can list all affixes that are combined with a list of stems. The role of unsupervised morpheme segmentation is to segment the words into stem and affix. This paper illustrates the importance of unsupervised morphological segmentation for the problem of morpheme boundary detection for resource poor languages which are highly inflectional and agglutinative in morphology. It specially focuses on few approaches which are based on their performance with highly agglutinative language.

Keywords*: *Hidden Markov Model (HMM), Morpheme, Morphological Segmentation, Suffix, Unsupervised Learning

I. INTRODUCTION

Morphology is the study of words and their inner structures [15] which aims to focus on the grouping of different word forms which diverges semantically and syntactically from the original word. It also enables the creation of new words irrespective of the language. Languages like Finnish, Turkish, Kannada, Tamil, and Bengali are agglutinative in nature because they have complex internal structures in which units join together to form new words. Basically words in agglutinative languages have a high number of morphemes. Morphological analyzer [1] plays a predominant role in preprocessing the morphological language to find the lemma and its morphological information. Initially the morphological process was tailored amidst the large amount of manual work done by linguistic expertise, but the need to discover a morphology automatically from annotated text, language are easier to achieve through the unsupervised machine learning. The traditional linguistic background work aims at morphological analysis, but morphological segmentation is an alternative that can be used as a full fledged morphological analysis. This paper mainly attributes the task of Morphological segmentation.

II. TASKS IN MORPHOLOGICAL LEARNING

NLP applications consist of three common tasks of morphological processing: segmenting the words, identification of word forms which are morphologically related, uncover the morphemes of the words using the morphological analysis (Refer Table1). This paper attempts to bring a broader view about the morphological segmentation which initiates to chunk the given words into morphemes

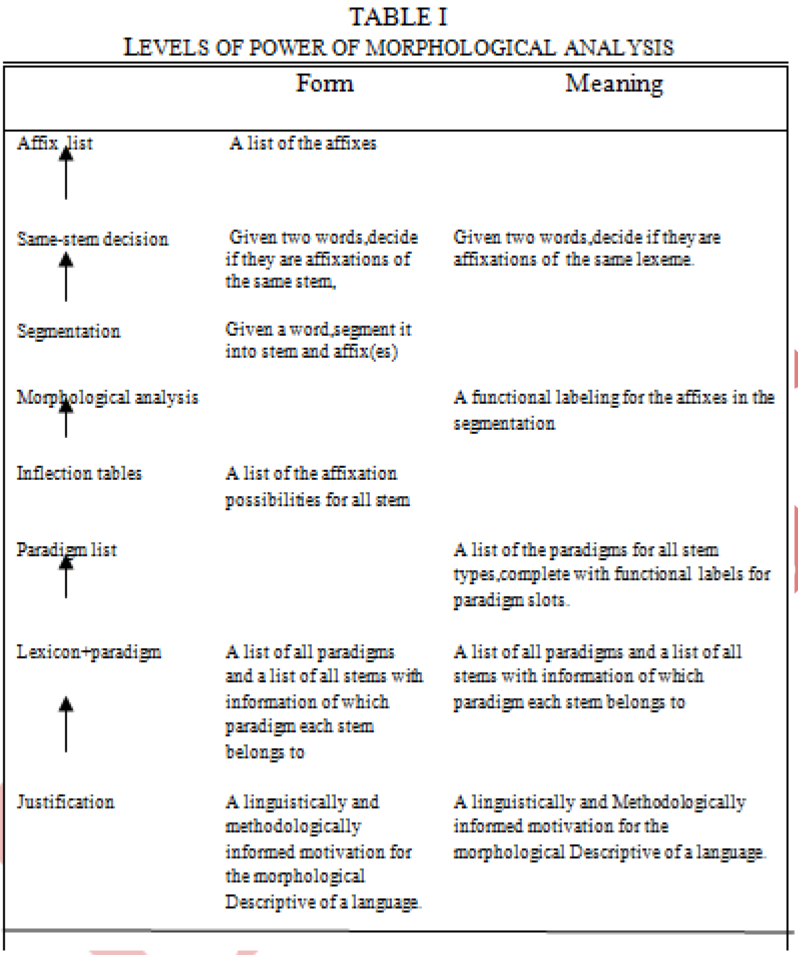

TABLE I
LEVELS OF POWER OF MORPHOLOGICAL ANALYSIS

	Form	Meaning
Affix list	A list of the affixes	
Same-stem decision	Given two words,decide if they are affixations of the same stem,	Given two words,decide if they are affixations of the same lexeme.
Segmentation	Given a word,segment it into stem and affix(es)	
Morphological analysis		A functional labeling for the affixes in the segmentation
Inflection tables	A list of the affixation possibilities for all stem	
Paradigm list		A list of the paradigms for all stem types,complete with functional labels for paradigm slots.
Lexicon+paradigm	A list of all paradigms and a list of all stems with information of which paradigm each stem belongs to	A list of all paradigms and a list of all stems with information of which paradigm each stem belongs to
Justification	A linguistically and methodologically informed motivation for the morphological Descriptive of a language.	A linguistically and Methodologically informed motivation for the morphological Descriptive of a language.

III. MORPHOLOGICAL SEGMENTATION

Morphological segmentation or word decomposition is a constructive approach for specific applications and languages. It is a process of analyzing a word by identifying its constituent morphemes. Speech recognition, machine translation and information retrieval rely on a constructive vocabulary and statistical language model for correct formation of words. Hereby segmenting a word poses an important challenge on morphological rich languages. Since the supervised learning or a rule based approach has chances to uncover many word forms it is hectic has it rely on linguistic expertise. Hence the evolution of unsupervised morpheme analysis leads to the process of automation that replaces the manual process role. For highly agglutinative language the set of morphs is not essentially different from the set of morphemes. Therefore the process of segmentation is straightforward since the output of the segmented algorithm is an ordered list substrings, morphs and the strings are same that are found from the reference segmentation.

4. APPLICATION AREAS OF MORPHOLOGICAL SEGMENTATION

4.1. Machine Translation

Morphological segmentation plays a predominant role in the field of machine translation. It handles morphological information in various phases of machine translation systems. Some machine translation approaches used morphological information within the preprocessing step [2] [16] [5] in which the translation process avails the morphological knowledge. Few models like factored models [4] [17] [18] [19] in machine translation systems utilize morphological segmentation to incorporate the additional knowledge about words. Invariably translation systems use morphological segmentation within post processing step [20] [21] where the stemmed texts act upon the translation and morphological form of words.

4.2. Speech Recognition

Speech recognition is one of the fields that exclusively depends on morphological segmentation.These systems are based on word dictionary along with language models in which the language model investigates the sequence of constituents in a particular language.For morphological rich languages like Turkish, Arabic, Kannada, Telugu, Tamil constructing a word dictionary is a challenging task. Language models are modeled with morpheme than words which resolves the out of vocabulary and problem of data sparsity. Morpheme modeling exhibits its usage in speech recognition systems for morphologically rich language like Finnish, Estonian, Turkish[6].For Turkish language recognition units are used instead of words[5].Several approaches[7] which constituents morpheme are used to deal with data sparsity of Arabic language

4.3. Information Retrieval

Information retrieval researchers use morphological segmentation to solve the ambiguity and the formation OOV words. For highly morphological rich languages handling this ambiguity and OOV words is an exigent task. Stem generation [8] Lemmatisation [9] are prominent approach that solves the problem of ambiguity of extracting base forms of words

V. METHODS IN UNSUPERVISED MORPHOLOGICAL SEGMENTATION

In Natural language processing community several unsupervised morphological analysis have been put into practice for English .In recent years the unsupervised methods are expanding to analyze morphological rich languages other than English. We focus our attention on three approaches [23] that are used to find the morphological factors of highly agglutinative language like Finnish Bengali, Kannada.

5.1. Linguistica

Linguistica is a tool that implements the technique of Goldsmith's method[10] of unsupervised learning of morphology which is based on the idea of minimum description length[11] (MDL).Generally MDL consists of four parts: a model of consists of the data assignd with a probability distribution from where the data is drawn,followed by the second model where a compressed length is assigned to the data.This is purely based on familaiar information theoretic notions.This proceeds to the model assigned with a length followed by the a model which handles the opimal analysis of data(i.e) the sum of the length of the compressed data and the length of the model is the smallest.In précis MDL is nothing but the combination of the length of morphology to the length of compressed data.For a given corpus of unannotated text a set of signatures are produced where the signature is a pattern consists of affixes that the stem can utilize to generate a word. For Example: the suffix signature in English could be NULL. Ed, in, is, combines with the stem laugh to create the words laugh, laughed, laughing, laughs. Therefore a list of stems, prefixes, suffixes and frequency information are provided by this algorithm.

5.2. Morfessor Categories-Map

Morfessor a language independent [12], data driven method for the unsupervised morphological segmentation is applied successfully for various languages. Minimum description length (MDL) and maximum a posterior (MAP) are the techniques implemented in Morfessor to optimize the accuracy with minimal model complexity to perform the task of segmentation. For segmenting the highly agglutinative language the current state of art Morfessor Categories a generative probability model [13] is used. This model segments the words in the corpus using Hidden Markov model (HMM) where the hidden states are latent morph categories. Prefix, Suffix, Stem and the additional non-morpheme category termed as noise are the described categories. In this model each morph lexicon consists of hierarchical entries. This feature of the algorithm supports the agglutinative word structure of complex words, since each morph can either be a string of letters or two submorphs, which can recursively hold a sequence of submorphs. It encapsulates a parameter (the perplexity threshold b) that manifest the optimal performance. This model attains an F-measure value of 70% for highly agglutinative languages like Turkish and Finnish.

5.3. Language-Independent Morphological Segmentation

This method is probably applied to unsupervised learning of morphological parsing of Indo-Aryan languages [14]. This algorithm is an extension of Keshava and Pitler's algorithm for morpheme induction. The main key idea behind the Keshava and Pitlers algorithm is to use words that appear as substrings of other words and transitional probabilities together to detect morpheme boundaries. [22]. The extension of this list is modified by employing a length dependent threshold that prunes the list of candidate affixes, detection of composite suffixes through the strength of the suffix and the word level similarity and it move ahead by inducing simple idea of relative corpus frequency of candidates. . The predominant feature of this algorithm is to move beyond one slot morphology by handling the words which have multiple suffixes and effective identification of inappropriate morphemes attached to it. Thus, this algorithm predicts an F-score of 83.29% on Bengali language.

VI. CONCLUSION

This paper describes the nature of unsupervised morphological segmentation which plays a prevalent role in the morphological analysis of language. The choice of algorithm for modeling the unsupervised morphological segmentation were based on the performance on morphological rich languages like Bengali, Finnish and Kannada. In future the Dravidian languages like Tamil, Telugu, Malayalam which are highly agglutinative in nature can be segmented through the described algorithms. Thus the methods elucidated can pave a way for morphological segmentation among the agglutinative language which can eventually benefit the morphological learning of languages.

REFERENCES

Journal Papers:

[1] Ananthi Sheshasaayee and Angela Deepa.V.R , "The Role of Morphological Analyzer and Generator for Tamil Language in Machine Translation Systems", International Journal of Computer Sciences and Engineering, Volume-02, Issue-05, Page No (107-111), May -2014

[2] Brown, Peter F., et al. "The mathematics of statistical machine translation: Parameter estimation." Computational linguistics 19.2 (1993): 263-311.

[3] De Gispert, Adriá, and José B. Mariño. "On the impact of morphology in English to Spanish statistical MT." Speech Communication 50.11 (2008): 1034-1046.

[4] Dr. Ananthi Sheshasaayee and Angela Deepa. V.R "The Transition of Phrase based to Factored based Translation for Tamil language in SMT Systems",International Journal of Engineering research and general science, Volume 2, Issue 4, June-July, 2014

[5] Arısoy, Ebru, Helin Dutağacı, and Levent M. Arslan. "A unified language model for large vocabulary continuous speech recognition of Turkish." Signal Processing 86.10 (2006): 2844-2862.

[6] Creutz, Mathias, et al. "Morph-based speech recognition and modeling of out-of-vocabulary words across languages." ACM Transactions on Speech and Language Processing (TSLP) 5.1 (2007): 3.

[7] Vergyri, Dimitra, et al. "Morphology-based language modeling for arabic speech recognition." INTERSPEECH. Vol. 4. 2004.

[8] Kettunen, Kimmo, Tuomas Kunttu, and Kalervo Järvelin. "To stem or lemmatize a highly inflectional language in a probabilistic IR environment?." Journal of Documentation 61.4 (2005): 476-496.

[9] Koskenniemi, Kimmo. "Two-Level Model for Morphological Analysis." IJCAI. Vol. 83. 1983.

[10] Goldsmith, John. "Unsupervised learning of the morphology of a natural language." Computational linguistics 27.2 (2001): 153-198.

[11] Rissanen, J. "Stochastic complexity in statistical inquiry, 1989." World Scientific, River Edge, NJ.

[12] Creutz, Mathias. Induction of the morphology of natural language:Unsupervised morpheme segmentation with application to automatic speech recognition. Helsinki University of Technology, 2006.

[13] Creutz, Mathias, and Krista Lagus. "Unsupervised models for morpheme segmentation and morphology learning." ACM Transactions on Speech and Language Processing (TSLP) 4.1 (2007): 3.

[14] Dasgupta, Sajib, and Vincent Ng. "Unsupervised morphological parsing of Bengali." Language Resources and Evaluation 40.3-4 (2006): 311-330.

Books:

[1] Bauer.L.(2003).Introducing Linguistic Morphology (3rd ed.).22 George Square,Edinburgh:Edinburgh University Press.

Proceedings Papers:

[1] Goldwater, Sharon, and David McClosky. "Improving statistical MT through morphological analysis." Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2005.

[2] Yang, Mei, and Katrin Kirchhoff. "Phrase-Based Backoff Models for Machine Translation of Highly Inflected Languages." EACL. 2006.

[3] Koehn, Philipp, and Hieu Hoang. "Factored Translation Models." EMNLP-CoNLL. 2007.

[4] Avramidis, Eleftherios, and Philipp Koehn. "Enriching Morphologically Poor Languages for Statistical Machine Translation." ACL. 2008.

[5] Minkov, Einat, Kristina Toutanova, and Hisami Suzuki. "Generating complex morphology for machine translation." ACL. Vol. 7. 2007.

[6] Toutanova, Kristina, Hisami Suzuki, and Achim Ruopp. "Applying Morphology Generation Models to Machine Translation." ACL. 2008.

[7] Keshava, Samarth, and Emily Pitler. "A simpler, intuitive approach to morpheme induction." Proceedings of 2nd Pascal Challenges Workshop. 2006

[8] Bhat, Suma. "Morpheme segmentation for kannada standing on the shoulder of giants." 24th International Conference on Computational Linguistics. 2012.

Biographical Notes

Dr.Ananthi Sheshasaayee working as Associate Professor at PG & Research Department of Computer Science ,Quaid-e-Millath Govt College for Women,Chennai.India

Ms.Angela Deepa.V.R currently pursuing Ph.D in Computer Science at PG & Research Department of Computer Science ,Quaid-e-Millath Govt College for Women,Chennai.India