Morpheme Segmentation for Highly Agglutinative Tamil Language by Means of Unsupervised Learning

Ananthi Sheshasaayee, Ph.D.
Research Supervisor
Department of Computer Science
QMGCW, University of Madras
Chennai-2

Angela Deepa V. R.
Research Scholar
Department of Computer Science
QMGCW, University of Madras
Chennai-2

ABSTRACT
To understand human language is one of the major challenges in the field of intelligent information systems. Morphological processing is the first step to be done in many Natural language processing applications. This task becomes crucial for morphological rich languages. This paper illustrates the importance of unsupervised morphological segmentation algorithms for the problem of morpheme boundary detection for Tamil language which are highly inflectional and agglutinative in morphology. This paper serves as ground work to represent the various methods and the comparative study among the selection of the algorithms which is based on highly agglutinative languages like Kannada, Finnish and Bengali. The prime advantages of these algorithms elevate to the efficient morphological processing of Tamil language

Keywords
Inflection, Morphological Segmentation, Natural language processing, Suffix, Unsupervised learning

1. INTRODUCTION
In most of the NLP applications, a morphological component termed as morphemes are the fundamental and basic unit that serves as a bridge between the texts and structured information of the language vocabulary. Languages are generally classified according to the morphological complexity. Morphological complexity refers to the degree to which the languages use bound morphemes, typically to express grammatical or derivational relations. Therefore the morphological system needs considerable amount of work to label the linguistic information to the morphological component manually. To overcome this machine learning approaches are used. Automated morphological analysis plays a predominant role in many natural language applications such as speech recognition, machine translation etc. These applications make use of words as vocabulary base units. Languages like Tamil. Telugu, Malayalam, Kannada, Finnish, Turkish, Bengali are termed as agglutinative in nature because the units join with one another to form new words which patterns a complex internal structures. Therefore the words present in the agglutinative languages can contain high number of morphemes.

Morphological analyzers [1] play a prime role in preprocessing the morphological language to ascertain the lemma and its morphological information. Initially the morphological process was carried out by linguistic expertise in which the texts were annotated manually. It caused a tedious tribulation as it relies on the human factor to meet their needs. This problem is resolved by the unsupervised machine learning. In the task of unsupervised morphological learning based on the statistical regularities in the input data the desired output is predicted.. The morphological segmentation serves as a ground work to understand the morphological component which can prevail in the process of an effectual morphological analysis of agglutinative languages. This paper mainly attributes the task of Morphological segmentation.

Morphological segmentation also termed as word decomposition is a productive approach for building specific applications in terms of language processing. It involves a process of scrutinizing a word by identifying its constituent morphemes. The main chore of morphological segmentation is to segment the given word forms or tokens into a set of morphs by identifying the each morpheme boundary location within the token. For highly agglutinative language the set of morphs is not intrinsically different from the set of morphemes. Hence the process of segmentation is undemanding which results in a full fledged analysis.

2. MORPHOLOGICAL SEGMENTATION METHODS
For highly agglutinative languages like Tamil the identification of morphemes attributes to various functionalities. This influences to the task of morphological segmentation. The segmenting of words into morphemes can be either done in supervised or unsupervised methods. The segmenting of words into morphemes can be either done in supervised or unsupervised methods. The following (Figure 1) describes the various methods involved in segmentation process.

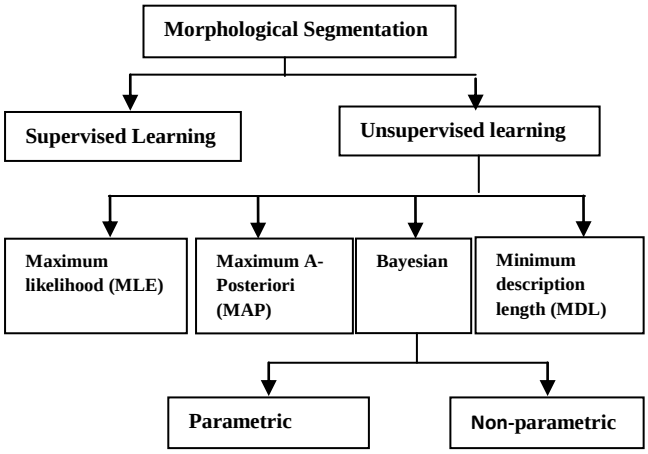

Figure.1 Morphological segmentation methods

2.1 Minimum Description Length (MDL)
This method is based on information theory which corresponds to the representation of morphology and resulting lexicon. Rissanen [2] proposed this method in which the reasoning was performed based on the ideology of

information theory.MDL outperform the best compression of data by searching the hypothesis. Based on the regularities in the data the compression of data is performed. The ratio of data learned relies on the performance of the compression [3].

For example:

Sample data set D= {laughed, knocked, laughs, knocks}

Compressed data= {laugh, knock, s, ed}

Thus the compressed data as all fundamental information about the actual data .MDL principle are used in various frameworks which can classified as frequentist or Bayesian approach..

2.2 Maximum Likelihood Estimate (MLE)

This method gives the highest probability of data under the given hypothesis. Likelihood is termed as the probability of data under the hypothesis stated. MLE [4] mainly focus on the good estimation of parameters on the grounds of the observed data. Since the model and the empirical distribution are analogous they have a minimum divergence, which leads to incorporate to the concept of minimization of the Kullback-Leibler (KL) divergence in them [5]. While fitting this credo in the mode of segmentation the hypothesis that outputs the scrutiny in which the stem is considered as the full word and the suffix is an empty character leads to MLE. The reason behind this mark is that the hypothesis does not split the words exactly as the empirical data which attribute the KL divergence in between the hypothesis and the empirical zero [6].MLE needs sufficient amount of data for decent estimation.

2.3 Maximum a Posteriori Estimate MAP

This extension of MLE method is the maximum a posteriori, which integrates a prior distribution over hypothesis. Since MLE does not involve prior distribution over hypothesis, the likelihood along with the prior probability which determines the posterior probability leads to the maximum a posteriori estimate (MAP) [7]. The prior information takes up two forms namely informative and non-informative. Informative prior gives expressive information about the each hypothesis, whereas the non informative prior does not provide important information about the stated hypothesis. The MLE and MAP fail to estimate the probability distribution over hypotheses.

2.4 Bayesian Modelling

Bayesian modeling is devised according to the ideas of enclosing probability distribution over the target instances. This model can be either parametric or a non parametric. This model is based on Bayes theorem [8].

2.4.1 Parametric: According to the Bayes theorem the inverse probability distribution over the parameters uses the likelihood and the prior probability. Posterior probability is states the feasible use of parametric values for the observed data with with that of the likelihood and prior probability. When the probability of data is computed along with the intended values of the parameters which is used or normalization is called marginal probability of data. Parametric values are can either discrete or continuous.

2.4.2 Non parametric: Non parametric means that there are a vast number of parameters which highly rely on the growth of the data. This framework is pragmatic and supple which attains to hold the irregularities in the data by allowing the flexibility in the parametric space.

3. CHALLENGES IN TAMIL LANGUAGE

Tamil is one of the longest surviving classical languages in the world. It is spoken by more than 66 million people all over the world. Tamil is a morphological rich content language and quite complex since it inflect to person, gender and number markings and also combine with the auxiliaries that indicate aspect, mood, causation, attitude etc. Noun root inflects with plurals, oblique, case, postpositions and clitics. Therefore a single noun form can inflect to more than five hundred word forms along with postpositions. A single verb root can inflect to more than two thousand word forms including the auxiliaries. Eventually the identified roots are to be tagged at the word level for further language processing. But due the complexity of the verb forms, capturing it in a machine analyzable and gene ratable form is a challenging task.

4. APPROACHES ON UNSUPERVISED MORPHOLOGICAL SEGMENTATION

Initially the practice of unsupervised methods has been used in English language. This influenced the other languages especially morphological rich one to use unsupervised learning for morphological analyzing.We focus our attention on three approaches that are used to find the morphological factors of highly agglutinative language like Finnish, Bengali and Kannada

4.1 Linguistica

Linguistica is a tool that features Goldsmith method[9] of unsupervised learning of morphology. It is centered with an idea of Minimum description length (MDL)[10].In general the MDL consists of four parts: a model with a probability distribution assigned to a data from where the data is drawn, followed by a second model to which a compressed length is assigned based on familiar information theoretic terms to the data. It ensues to a model assigned with the length proceeds with the model that handles optimal analysis of data. The concept of MDL is condensed as a permutation of the length of morphology to the length of the compressed data. The unannotated texts present in a given corpus produce signatures which are in a pattern incorporating the affixes that the stem or the root word uses to form a word.

For Example: the suffix signature in English could be NULL, ed, ing, s, that combines with the stem **dance** to create the words like dance, danced, dancing, dances. Thus this algorithm predicts the list of stems, prefixes, suffixes and frequency information effectively.

4.2 Morfessor Categories-MAP

Morfessor is an unsupervised morphological segmentation which implements the data driven method and are language independent [11] in nature. To perform the task of segmentation it optimizes the accuracy of the minimum model complexity with the help of techniques like minimum description length (MDL) and maximum a posterior (MAP).Morfessor Categories, a generative[12] probabilities model is the current state of art used for segmenting highly agglutinative languages .In this model words of the given corpus is segmented using Hidden Markov Model(HMM).The hidden states in this model are stated as latent morphs categories and the described categories or noise consists of stem,prefix,suffix and the additional non-morpheme categories.

This model structures the morph lexicon as hierarchical entities which in turn benefits the process of segmenting

highly agglutinative word structures which are in complex form.Each morph in a agglutinative word structure can contain two or more submorphs which are recursively sequenced to hold the submorphs. It consist of a parameter (the perplexity threshold *b*) that sets the optimal performance of this algorithm.For highly agglutinative language like Turkish and Finnish this model attains an F-measure value of 70%.

4.3 Language-Independent Morphological Segmentation

This algorithm is applied for the unsupervised learning of morphological parsing of Indo-Aryan languages[13].The centre idea of morpheme induction uses the heuristics of Keshava and Pitler's algorithm.This algorithm is referred as UnDivide based on the line up complementary the publications[14].The core idea behind this algorithm is to make use of words that emerge out as substrings of the other words which integrates the transitional probabilities to detect morpheme boundaries[15]. In addition its employs a length dependent threshold that prunes the list of candidate affixes,and recognizes composite suffixes through the strength of the suffix and word level similarity.It encompasses relative corpus frequency of candidates for effective root induction. The noticeable feature of this algorithm is to move beyond one slot morphology to handle words which constituents multiple suffixes. It is efficient in identifying inappropriate morphemes attached word forms. This algorithm can manifest an F-score of 83.29% on Bengali language.

5. COMPARISON OF APPROACHES

The comparison of these approaches is based on the performance on agglutinative language Kannada [15]. The parameters used in this comparison are based on the data size, Morpheme boundary detection, and Specific nature of the algorithms that efficiently identify the morphemes.

5.1 Data Size

Data size plays a vital role in the morphological processing. Efficiency of the algorithms is predicted based on the size of the dataset. Among the three algorithms, Undivide algorithm outperforms best on a larger data set where the Morfessor-CatMAP is excellent on a small data set. On large data set Undivide algorithm shows better results on inflected nouns whereas the Morfessor-CatMAP shows reliable result in smaller data set than linguistica.While evaluating the inflected verbs on data sets irrespective of the size the Undivide algorithms outperform the Linguistica and Morfessor-CatMAP.

5.2 Morpheme Boundary Detection

Linguistica only separates the final affix in case of plural endings, Morfessor overly segment the words. But the ability of Morfessor to deal with complex structured language is high when compared to other algorithms.

5.3 Nature of the Algorithms

5.3.1 Linguistica: This algorithm does not support the compound words and its complexity in deriving the verbal inflectional system. The single slot capability of Linguistica is the stumbling block of the model.

5.3.2 Undivide algorithm: This algorithm is successful in generating the character-change rules by a single replacement. It allows the addition and deletion of morpheme boundaries elaborates and analyzes the derivational affixes. It would be reasonable if it finds out even the morphophonemic (internal sandhi) rules of languages.

5.3.3 Morfessor-CatMAP: This algorithm identifies the affixes effectively. In particular, it labels the prefixes and suffixes competently irrespective of their data size. It is quite successful in the derivational morphology

6. CONCLUSION

The unsupervised morpheme analysis for morphological complex language is a noticeable approach in the field of Language technology. Morphological segmentation of words is important for morphological complex languages like Tamil which are agglutinative in nature since the amount of the word forms are based on the inflection, derivation and composition. The main aim of designing unsupervised morphological segmentation algorithms is to discover the morphemes which are suitable for most of the NLP tasks like large vocabulary speech recognition (LVCSR), statistical machine translation (SMT) and information retrieval (IR). This paper presents an elaborate view on unsupervised morphological segmentation and a comparative study of various segmentation algorithms. The approaches prescribed in this paper showed good results for Kannada language. Since the Kannada language and Tamil belongs to the same Dravidian family the best attributed approach can be used for Tamil language as well. This can eventually lead to the successful creation of morphological segmentation approach which can eventually lead to the better morphological analysis of Tamil word forms.

7. REFERENCES

[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] Rissanen, Jorma. "Modeling by shortest data description." Automatica 14.5 (1978): 465-471.

[3] Grünwald, Peter. "A tutorial introduction to the minimum description length principle." (2005).

[4] Myung, In Jae. "Tutorial on maximum likelihood estimation." *Journal of mathematical Psychology* 47.1 (2003): 90-100.

[5] Kullback, Solomon, and Richard A. Leibler. "On information and sufficiency."*The Annals of Mathematical Statistics* (1951): 79-86.

[6] Goldwater, Sharon J. "Nonparametric Bayesian models of lexical acquisition." PhD diss., Brown University, 2007.

[7] Gauvain, Jean-Luc, and Chin-Hui Lee. "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains." *Speech and audio processing, ieee transactions on* 2.2 (1994): 291-298..

[8] Stigler, Stephen M. "Who discovered Bayes's theorem?." *The American Statistician* 37.4a (1983): 29

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

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

[11] Creutz, Mathias. Induction of the morphology of natural

language: Unsupervised morpheme segmentation with application to automatic speech recognition. Helsinki University of Technology, 2006.

[12] 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.

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

[14] Dasgupta, Sajib, and Vincent Ng. "High-Performance, Language-Independent Morphological Segmentation." *HLT-NAACL*. 2007.

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

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