MORPHEME BOUNDARY IDENTIFICATION USING LETTER SUCCESSOR VARIETY

Indu Joseph Thoppil*[1], Elizabeth Sherly[2]

[1]Indian Institute of Information Technology and Management-Kerala
Indujoseph.mphilcs2@iiitmk.ac.in[1]
[2] Indian Institute of Information Technology and Management-Kerala
sherly@iiitmk.ac.in[2]

Abstract: Morpheme boundary identification is one of the prominent problems in morphological analysis of NLP applications. This study proposes an alternative technique to effectively identify the boundary of each morpheme from a compound word. This is useful in wide range of NLP applications from stemming, word assistance to document categorizers. Malayalam, a major South Indian language has the linguistic capability to have high number of morphemes per word. The present study sets out to discover the effectiveness of Letter Successor Variety techniques could be useful for identifying the morpheme boundary in Malayalam. Letter Successor Variety is based on statistical co-occurrence measures and contextually similar words.

INTRODUCTION

Computer systems with a level of understanding of human language are a requirement of the today's needs. Researchers are very much interested in developing systems to improve the interaction between humans and computers. In order to achieve it, research is increasingly applied to automated processing of natural language data. This is useful for many applications such as information retrieval, text categorization, dictionary automation, text compression, data encryption, vowelization and spelling aids, automatic translation and computer aided instruction.

According to standard linguistic theory, words consist of morphemes, which are the smallest individually meaningful elements in a language. Since an immense number of word forms can be constructed by combining a limited set of morphemes, the capability of understanding and producing new word forms depends on knowing which morphemes are involved(e.g., water, water +s , water +y, water +less, water +less + ness , sea +water).Morpheme boundaries are not normally marked in text unless they coincide with word boundaries.

In general, methods for automatic morpheme boundary identification can be divided into three main categories.

- ***Rule based approaches*** which identified morphemes by using predefined rules for word segmentation in a given language.
- ***Data driven approaches*** which are based on statistical techniques.
-
- ***Hybrid approaches*** which combine both mentioned above.

One of the first data-driven approaches is the *semantically oriented segmentation of words*, proposed for the inflected languages that are based on the principle of the *latent semantic analysis*, while boundaries for word segmentation are given by the branching factor in a tree structure for a group of words with similar properties (affix candidates)**.**

Among the most well-known word segmentation methods is the *letter successor variety* (LSV) [6] *segmentation*. This knowledge-free morphology segmentation is the base for the other different data-driven methods. Algorithm is based on the computing frequency of distribution of the character variants after (or before) the group of characters (respectively) in a given word. The segmentation boundaries are specified in the places after (or before) maximal (or minimal) value of the occurrence variety of characters on the given position.

Similar approach, called the minimum description length (MDL) principle was proposed for Finnish [4]. This relatively complex unsupervised data-driven algorithm considers a model that would be able to describe examined language, the morphological regularities in a language or entire model by using the probability distribution of morphs in a group of words with similar properties in the simplest and shortest way. MDL algorithm was later successfully applied in the system called Morfessor for the modeling of the Finnish language using statistical morphs [5], later in modeling Slovenian [4], agglutinative Estonian, inflective

Turkish or conversational Arabic using morph-based LMs [3].

One of the rule based approach is the *suffix stripping approach*. Once the suffix is identified, the stem of the whole word can be obtained by removing that suffix and applying proper orthographic (sandhi) rules. A set of dictionaries like stem dictionary, suffix dictionary and also using morphotactics and sandhi rules, a suffix stripping algorithm successfully implements the morpheme boundary identification.

Also *stemmer based approaches* uses a set of rules containing list of stems and replacement rules to stripping of affixes. It is a program oriented approach where developer has to specify all possible affixes with replacement rules. Potter algorithm is one of the most widely used stemmer algorithm and it is freely available. The advantage of stemmer algorithm is that it is very suitable to highly agglutinative languages like Dravidian languages for identifying the morpheme boundary.

Malayalam one of the Dravidian languages, is a highly agglutinative, inflectional and a relatively free word order language. Considering all these features, identification and segmentation of words into morphemes is a challenging task. In order to address the problem, we suggest a new boundary detection technique using Letter Successor Variety (LSV)[1][2]. The algorithm detects morpheme boundaries and can also be modified to perform other tasks, e.g. clustering of word forms of the same lemma and the classification of the found morphemes. The primary aim is to reach maximum precision, so that the output can be used in a post processing machine learning step.

METHODOLOGY AND IMPLIMENTATION

In Malayalam, common approaches used for morphological analysis are Suffix stripping method [6][8] and Paradigm method. Since Malayalam word consists of lengthy sequences of stems and suffixes, they cannot handle the possibly high number of morphemes per word. The Letter successor Variety technique measures the amount of various letters occurring after a given substring with respect to some context of other words (in this case semantically similar ones), weighting that value according to bi and trigram probabilities and comparing the resulting score to a threshold. Hence it is designed to handle agglutinative morphologically rich language. This is done by computing LSV score. The LSV based technique computes several values for each transition between characters of a given word form: the left and the right letter successor variety, the overlap factor, the inverted bigram score.

Computing the values for LSV based techniques is done as follows:

•**Letter successor variety**: Count the number of different letters encountered after a given string from the beginning (or the end, respectively).

•**Overlap factor**: In order to detect that a smaller string is part of a larger morpheme, it is possible to count how many of the words seen with a particular suffix have also been seen with another substring.

•**Inverted Bigram score**: uncertainty weight has temporarily been introduced for normalization purpose.

The final LSV score for any transition n can then be computed by the multiplication of the initially obtained LSV with the averaged overlap factor and the inverse bigram weight. A high score translates into a morpheme boundary by means of a threshold. There are various possibilities to interpret such scores. In this first prototype, a simple threshold has been introduced, as another free parameter. All final scores above the threshold are considered to mark morpheme boundaries and the words are then segmented using these boundaries.

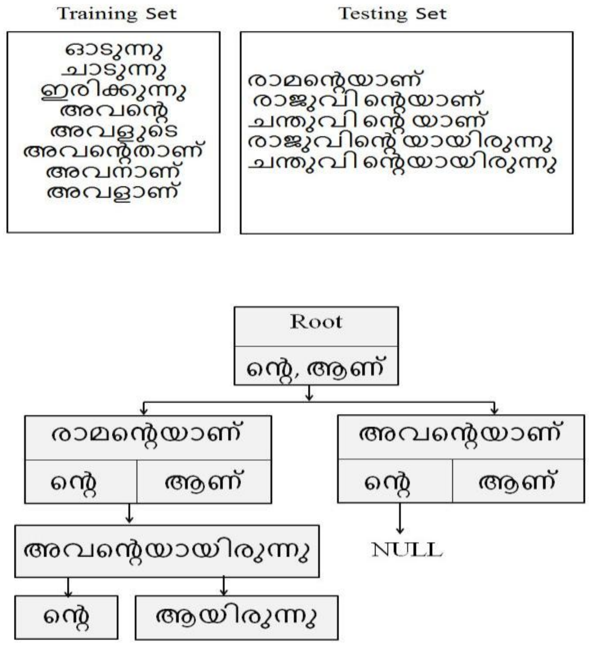
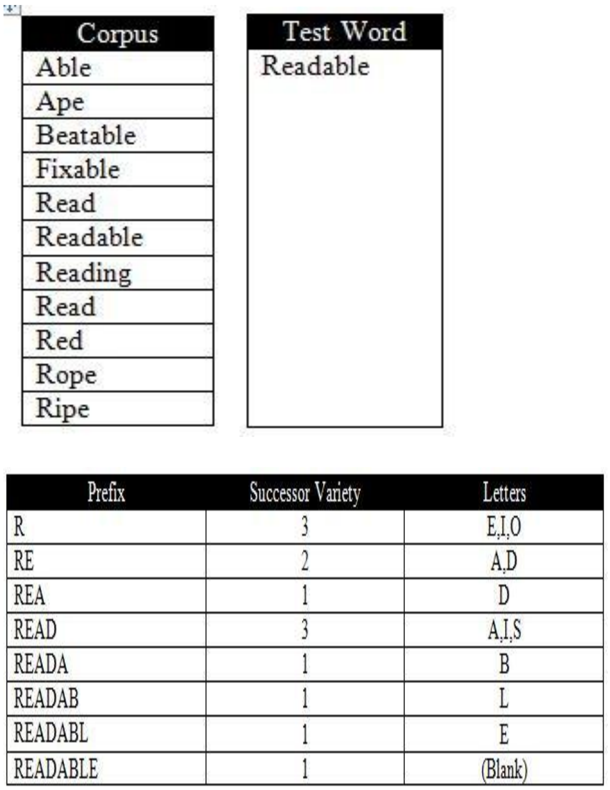


Figure 1: Illustration of morpheme boundary identification using LSV



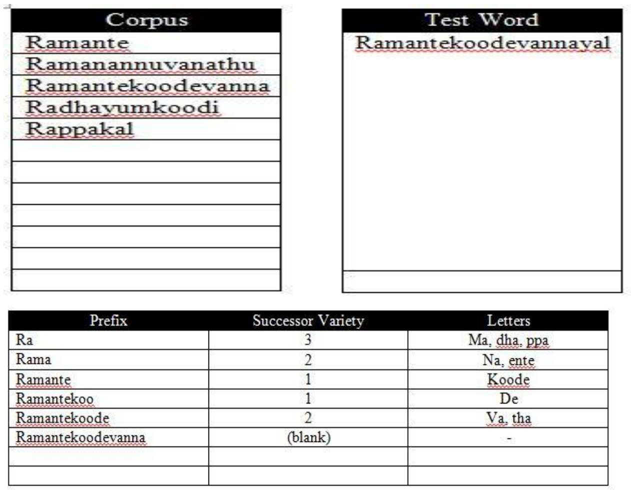
Corpus
Able
Ape
Beatable
Fixable
Read
Readable
Reading
Read
Red
Rope
Ripe

Test Word
Readable

Prefix	Successor Variety	Letters
R	3	E,I,O
RE	2	A,D
REA	1	D
READ	3	A,I,S
READA	1	B
READAB	1	L
READABL	1	E
READABLE	1	(Blank)

Figure 2: LSV stem process for English words

Malayalam words. When this process is carried out using a large body of text, successor variety of substrings of a term will decrease as more characters are added until a segment boundary is reached.



Corpus
Ramante
Ramanannuvanathu
Ramantekoodevanna
Radhayumkoodi
Rappakal

Test Word
Ramantekoodevannayal

Prefix	Successor Variety	Letters
Ra	3	Ma, dha, ppa
Rama	2	Na, ente
Ramante	1	Koode
Ramantekoo	1	De
Ramantekoode	2	Va, tha
Ramantekoodevanna	(blank)	-

Fig3: LSV stem process for English words

The formula to calculate the Boundary Identification is given below

Lsv (w, I) = plsv (w, I) fw(w, I) ib(w, I): (1)

- plsv (w, I) is the plain number of different letters found to the right of the substring between the beginning of the word w and the position i.
- fw(w, I) is the bi or trigram based frequency weight of the substring.
- ib (w, I) is the inverse bigram weight used entirely for normalization purpose

CONCLUSION

This study shows the potential for significant improvement in morpheme boundary identification for a complex morphological language like Malayalam. Our technique takes advantage of contextually similar word forms. The result is highly useful for morphological analysis and has a high impact value in the field of Information Retrieval for stemming purposes.

FORTHCOMING RESEARCH

A performance boost can be achieved by combining the machine learning algorithms. This will be a significant step towards a true morphological segmentation similar to what can be done manually. Currently LSV examines words in isolation. If word sequences were utilized instead, larger idiomatic segments could be discovered, such as multi-word geographical names: San Francisco, New York City, New Zealand.

REFERENCE

[1] ROTOVNIK, T., M. S. MAUCEC and Z. KACIC. Large Vocabulary Continuous Speech Recognition of an Inflected Language using Stems and Endings. *Speech Communications*. 2011, vol. 49, iss. 6, pp. 437-452. ISSN 0167-6393. DOI: 10.1016/j.specom.2007.02.010.

[2] KARPOV, A., I. S. KIPYATKOVA and A. RONZHIN. Very large Vocabulary ASR for Spoken Russian with Syntactic and Morpheme Analysis. In: *Proceedings of INTERSPEECH'2011*. Florence: ISCA, 2011, pp. 3161-3164. ISBN 978-1-61839-270-1.

[3] SCHONE, P. and D. JURAFSKY. Knowledge-Free Induction of Morphology using Latent Semantic Analysis. In: *Proc. of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language*

Learning. Lisbon: Association for Computational Linguistics, 2000, pp. 67-72. DOI: 10.3115/1117601.1117615.

[4] HARRIS, Zellig S. From Morpheme to Phoneme. *Language*. 1955, vol. 31, no. 2, pp. 190-222. ISSN 0097-8507.

[5] BORDAG, Stefan. Unsupervised Knowledge-Free Morpheme Boundary Detection. In: *Proc. of International Conference on Recent Advances in Natural Language Processing, RANLP'2005*. Borovets: INCOMA, 2005, pp. 1-7. ISBN 95491743-3-6. DOI: 10.1.1.109.2571.

[6]Bordag, S: Two-step approach to unsupervised morpheme segmentation. In: Proceedings of the PASCAL Challenges Workshop on Unsupervised Segmentation of Words into Morphemes, Venice,Italy (April 2006).

[7] CREUTZ, M. and K. LAGUS. Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora using Morfessor 1.0. *Computer and Information Science*. 2005, vol. Tech. report A81, pp 27. ISSN 1913-8997.

[8] Bordag, S: Unsupervised knowledge-free morpheme boundary detection. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), Borovets,Bulgaria.(Sept 2005).

[9] Riyad Al-Shalabi , Ghassan Kannan , Iyad Hilat , Ahmad Ababneh and Ahmad Al-Zubi , 2005. Experiments with the Successor Variety Algorithm Using the Cutoff and Entropy Methods. *Information Technology Journal, 4: 55-62*.

[10] Prajitha U, Sreejith C, P.C. Reghu Raj ," LALITHA: A Light Weight Malayalam Stemmer Using Suffix Stripping Method ", IEEE International Conference on Control Communication and Computing (ICCC) 13-15 December 2013 Trivandrum.

[11] Rinju O.R., Rajeev R. R., Reghu Raj P.C., Elizabeth Sherly, " Morphological Analyzer for Malayalam: Probabilistic Method Vs Rule Based Method" , IJCLNLP volume 2 issue 10 October 2013.