Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 8
Open Access

Exploring Morpheme Length and Classification Accuracy in Dravidian Languages: An Unsupervised Learning Approach

Thivaharan.S[1], Srivatsun.G[2]

[1]Assistant Professor (Sr. Gr), CSE Dept, KPR Institute of Engineering and Technology, Coimbatore
[2]Associate Professor, ECE Dept, PSG College of Technology, Coimbatore
[1]thivaharan.s@kpriet.ac.in, [2]gsn.ece@psgtech.ac.in

Cite this paper as: Thivaharan.S, Srivatsun.G (2024) Exploring Morpheme Length and Classification Accuracy in Dravidian Languages: An Unsupervised Learning Approach. *Frontiers in Health Informa 3987-3998*

ABSTRACT
Agglutination in south Indian languages like Tamil, Telugu and Malayalam enriches the volume of lexicons and vocabulary, resulting in a volume of capable morphemes. Though agglutination is considered as a prominent feature that enriches a language, it also leads to lesser accuracy in feature classification during morpheme handling and analysis, which eventually results in inappropriate mapping across languages of similar nature. Agglutination along with regional dialects, poses an open challenge in attaining high accuracy. In this paper, an unsupervised learning model based on polynomial regression framework is proposed for morphological segmentation and a study on how the morpheme length affects the classification accuracy is done. The model is based on unigram word segmentation with an assumption that morph length in the investigative data is evenly distributed. Two Morpho-tactically related language components, Informal Tamil and Deutsch (German language) were taken for consideration. Experimental results are benchmarked against the unique statistical morphological toolkit. The paper concludes through the experimental results claiming that morpheme length has definitive impact in the analysis and in improvising the prediction accuracy as close to 87%.

Keywords: Dravidian languages, Agglutination, Dialects, Morph length, polynomial regression.

1. INTRODUCTION

One of the crucial tasks on any natural language processing system is to handle and analyze the morphemes of the underlying language. Languages like Tamil (Informal) and German are inherently rich in inflections. Retrieval of information from a morphologically rich agglutinative language needs several reference layers which act as models for directing the prediction accuracy. Off-late along with the classic morphological analysis, morphological segmentation plays an additional role for enhancing the decision of morpheme validness.

Morphological segmentation splits the token in the word form in to meaningful morphemes, while setting a fixed justifiable boundary to the word under investigation. All the languages, even the low resourced one need the benefit of segmentation. Segmenting agglutinative languages with inflectional Morpho-tactic needs pre-trained models like MorphAGram.

MorphAGram as proposed by Ekshander et al [1], makes use of the linguistic components like context free grammar and regular expression (RegEx) [2]. MorphAGram extends the CFG by associating probabilistic weight to the generative structure of the rules. Probabilistic approach not only establishes a bonding between the segmented morphemes but also distributes the generative structure of the grammar. Generally this approach is termed as generative grammars.

The appealing nature of multi-lingual cross linked framework structure of unsupervised approach is taken as the key factor for choosing the same as rightly analyzed by Talukdar et al [3], The outcome of the un-supervised approach rightly fits to the parametric need of Bayesian approach, thus yielding to maximal attainment in the prediction accuracy during morpheme segmentation [4].

Owing to the context sensitivity and the ability to infer the subjective morphemes from the general or surface level segments, bayesian approach is used considering the level of existing works in the research community towards informal Tamil contents and its allied associates like Deutsch. Languages with varying dialects (colloquial Tamil) [5] fall under Low Lexicon Level (LLL) languages. Bayesian approach is much suitable for such category of languages.

2. EXISTING WORK

In this section, a study over the existing proposal and analysis is made with the standard work accomplishments. This literature is mainly focused on morpheme length in the segmentation phase.

Micheal Hahn et al [6], 2022, explored Languages use inflectional morphology to convey grammatical information through morpheme strings. This study suggests that cross-linguistic morphological fusion, where one morpheme expresses multiple features, can be explained by optimizing processing efficiency, specifically the memory-surprisal tradeoff like fusing highly informative neighboring morphemes improving the processing efficiency. Using datasets from languages like Tamil, Telugu, Malayalam and Kannada, the study shows that fusion levels align with optimal morpheme ordering for efficiency and predict typological patterns including suppletion. The research also finds that both detailed quantitative measures and broad linguistic generalizations can be understood through optimizing predictability and memory complexity tradeoffs. Features highly correlated in usage are more likely to fuse, supporting the idea that related meanings are expressed together, and suggesting that language typologies arise from efficient cognitive processing.

Ronald Rosenfield et al 2020 [7], proposed an efficient exponential language model that treats entire sentences as single units, using features as their properties. This approach avoids the chain rule and is more effective for modeling global sentential phenomena over the existing models. The model's training involves sampling from an exponential distribution and addresses the challenges using Monte Carlo Markov Chain techniques resulting in smoothing and step-size selection. The author also presented a new feature selection method based on model-corpus discrepancies. The models effectiveness is demonstrated through competitive incorporating lexical and syntactic information.

Soroush Vosoughi et al, 2023 [8], introduced an automatic system for detecting Child-Directed Speech (CDS) in the context of studying child language development. The intention of this research is to make use ambiguous pronunciation of the child and aiming to improve the same for achieving scaled up accuracy. CDS is speech directed by caregivers towards infants, often found in corpora used for these studies alongside non-CDS speech. Manual annotation of CDS in large corpora becomes impractical, so the automatic CDS detector proposed in the paper addresses this challenge. The focus is on proposing and evaluating different sets of features for CDS detection including acoustic, linguistic, and contextual features. Using a combination of these features, the CDS detector achieves an accuracy of 88% and an F1 score of 87% using Naïve-Bayes classifiers.

G.Dias et al, 2021 [9], discussed the importance of Tamil language morphology and semantic analysis in natural language understanding. It introduces various methods for morphological and semantic analysis such as Finite State Transducers, Tree Adjoining Grammars and Support Vector Machines. Semantic analysis focuses on word similarity and is crucial in areas like Natural Language Processing

and Information Retrieval. The Universal Networking Language (UNL) is highlighted as a structured ontological graph-based representation of natural language aiding in information extraction-cum-retrieval due to its simplicity. The review methodology includes collecting research articles from various sources and filtering based on keywords related to UNL, semantic analysis, and morphological analysis. The text also provides a comprehensive overview of UNL, discussing its graph generation, models for morphological analysis, and semantic analysis systems based on Word Embedding and Word Overlapping models. Challenges in developing morphological analyzers, semantic analysis systems, and UNL are discussed.

K. Manoharan et al, 2022 [10], introduced a pre-balanced standard lexicon, achieving high accuracy rates in orthographic syllabification and grapheme to phoneme conversion tasks. MIphon employs the Automatic Speech Recognition (ASR) for the direct evaluation for Malayalam Language without compromising originality in the foundation of the language. The authors present performance analysis metrics such as computation time for lexicon creation and word error rate (WER) comparing MIphon with other automated tools for lexicon creation.

Based on the literature survey, it is evident that the existing works give importance the sentence as a whole with inclination towards Maximum Entropy (ME) [11] models compromising the sampling size, step size and smoothing the word boundary. The clear differentiation in terms of boundary between the existing model and the trained model is not highlighted. For informal aspects especially in the Dravidian languages need the semantic coherence for the various Morpho-forms like verb distribution, tense match and dialog order in the paragraph.

3. IMPACT OF SYNTACTICAL COHERENCE AMONG THE DRAVIDIAN LANGUAGES

Due to the existence of syntactical differences between the genre-rooted languages, all the morphological segmentation techniques irrespective of their nature perform a post-appending process. Most of the time, as claimed by Klavans, 2018 [12], this affixation process result in the extraction of single root morpheme. Recent days, modern segmentation techniques deploy a novelty by adding the synchronical and diachronical gaps to retain the reasoning continuum among the languages. Adapter grammar with their ability to fit and mark the boundary to the symbol-wise clarity is well suited and is been used mostly.

Most of the languages can be either classified under "Single Root Languages (SRL)" and "Coherent Languages (CL)", where single root languages need no affixation and they are fully transferrable using finite state transducers (FST), the coherent languages has the indefiniteness with the morpheme boundary and they are agglutinative and fusional in nature. Unlike SRLs the CLs have no one-to-one bonding between the lexical form and surface form. Also the CLs have the inherent generative capability, which by the way after informalizing with the regional dialects makes the prevailing rules to look unsuitable at some times. In this paper, the following two languages are considered:

German (Deutsch): Fusional, more generative
ex: Haup means capital, Stadt means state, they fuse together and become Haupstadt

Tamil (Informal/regional dialect specific): Agglutinative, more generative
ex: Avanumma? means Is he too?. This Avanumma can be splitted as per grammar in to Avan + Um + A.

Symbolically the word segmentation can be represented as given below omitting the formal or informal nature of the language:

$$P(B|W) \propto P(W|B)\, P(B) \quad (1)$$

Where:

P(B|W) is the probability of a boundary given a word.
P(W|B) is the probability of a word given a boundary.
P(B) is the prior probability of a boundary.

The goal is to maximize P(B|W) for each position in a word W considering the boundary B.
As per the proposal in this paper, the unsupervised Morpho-tactics based segmentation, considering an agglutinative word (X) as $X_1X_2X_3...X_n$

$$P(X_i \mid H) = N * M + c * P(M) / N * M \quad (2)$$

Where, where "M" is the expected morpheme from the word under investigation, "N" refers to the number of times the morpheme is referred in the corpus history, "c" refers to the word boundary inclusion constant. The probability for word boundary detection at character position "i" is

I (Xi == Xj), {1, if both the morpheme extraction position of the word are at the same position
0, otherwise

As per the hypothesis probability, either 1 or 0 is produced as outcome, meaning 1 implies matching word boundary is identified and 0 implies no matching word boundary. The above probabilistic hypothesis base implication is repeated for the entire corpus of words.

4. FRAMEWORK

Considering the non-parametric nature of the informal Tamil and dialectal German language, the Bayesian model is tried for the corpus. The grammar to fix the lexical-rule correspondence is generally termed as the Adaptive Grammars [13] (AG) and they are context sensitive, making them highly suitable for boundary marker based morpheme segmentation models. The adapter grammar is based on the posterior probability which in turn solely rely upon the parsed sub tree, the resultant of the Morpho-tactics analogy. And as necessitated by the process, Monte Carlo snippets are used to back up the decision made by the model. This entire process is purely based on the un-adulterated content which is highly context sensitive. Based on the MorphAGram architecture the following phases are investigated and tried for this paper.

4.1. Grammar declaration

Context sensitive grammars need the decisive Terminals (T) and Non-Terminals (NT) along with their derivatives. CSG is a 4-tuple denoted as G=(N,Σ,P,S), where N is finite set of Non-Terminals, Σ is finite set of Terminals, S is the set of start symbols and P is finite set of production rules of the form

$$\alpha A\beta \rightarrow \alpha\gamma\beta, \text{ where } A \in N, \alpha, \beta \text{ and } |\gamma| \geq 1 \quad (3)$$

Context sensitive grammars impose equality in the left and right context marker length, thus resulting in length non-decreasing factorization. CSG treats a word as stemmed contextual form. When the complexity of the word manifolds, the non-terminals mitigates the elongation in the prediction process. Figure-1 illustrates the generated parse tree for the term. The language grammar associated with it is shown

S → NP Conj Emph
NP → Root Nom
Root → ava
Nom → n
Conj → um
Emph → ma

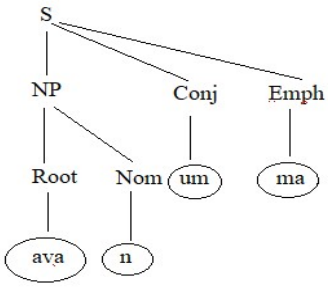

Figure 1 – Generated Parse for “avanumma”

4.2). Modeling the morpheme length extraction

The idea is to represent our assumptions or beliefs about the possible lengths of morphological units (morphemes) using a mathematical distribution. P_0 is the foundational distribution that expresses these beliefs. The Poisson distribution is selected as the model for morph lengths because it is suitable for representing the probability of a given number of events (morphs of a certain length) occurring within a fixed interval. The formula for the Poisson distribution [16] is

$$P(l, k) = l^k e^{-l} / k! \quad (4)$$

l denotes the lambda and its expected (average) morpheme length, k is the actual observed length and the formula calculates the probability of a morph having a specific length k given the expected length l. Two Base Distributions namely $P_0^A(m)$ and $P_0^B(m)$ directly uses the Poisson distribution to represent the probability of morph lengths also extends the Poisson distribution by incorporating an additional factor p(m), which represents some prior knowledge or additional probability associated with the morph m.

$$P_0^A(m) = P(l, k) = l^k e^{-l} / k! \quad (5)$$
$$P_0^B(m) = P(m) * P_0^A(m) \quad (6)$$

5. Training the model for morph length convergence

The learner requires two primary inputs: the grammar with adaptation information, and the vocabulary of the target language for segmentation. A unique unsegmented lexicon is termed as the vocabulary. If the vocabulary is quite large (e.g., over 50,000 words), it is advisable to provide only the most frequent words from the language. This allows us to inductively learn the segmentation of the remaining words. The scholar seeded approach, a list of morphemes. When some linguistic knowledge, such as a list of lexemes, is available, it can be integrated into the production rules of the resultant parsing tree.

5.1. Inference

The process of employing posterior and the estimates in Markov Monte Carlo involves inferring all the hyper parameters of the model, including the probabilities in the Probabilistic Context-Free Grammar [14] (PCFG) base distribution and the hyper parameters of the Pitman-Yor process. Figure-2 depicts the pitman-yor equivalent context free grammars

```
1 1 Word --> Prefix Stem Suffix
Prefix --> ^^^
Prefix --> ^^^ PrefixMorphs
1 1 PrefixMorphs --> PrefixMorph PrefixMorphs
1 1 PrefixMorphs --> PrefixMorph
PrefixMorph --> SubMorphs
Stem --> SubMorphs
Suffix --> $$$
Suffix --> SuffixMorphs $$$
1 1 SuffixMorphs --> SuffixMorph SuffixMorphs
1 1 SuffixMorphs --> SuffixMorph
SuffixMorph --> SubMorphs
1 1 SubMorphs --> SubMorph SubMorphs
1 1 SubMorphs --> SubMorph
SubMorph --> Chars
1 1 Chars --> Char
1 1 Chars --> Char Chars
```

Figure 2 – Pitman Yor context Free Grammar

5.2. Experimental Setup

Gibbs sampling as proposed by Yamakuchi et al 2022 [15], the current state of all variable draw repeatedly the equities framing the boundary ensuring the proper positioning of gap markers and character location. The corpus and the lexicon volume both morphologically segment the sampling sufficiency in term of frequency and numbers. The intuitive idea is that by sampling a sufficient number of times draws segmented morphological sequence of words in the entire process converges to the probabilistic distribution of the segmented words of the entire corpus. Algorithm 1 provides a general outline of how the Gibbs sampling procedure is applied to morphological segmentation. Figure-3 details the flowchart for the proposed Gibbs sampling approach.

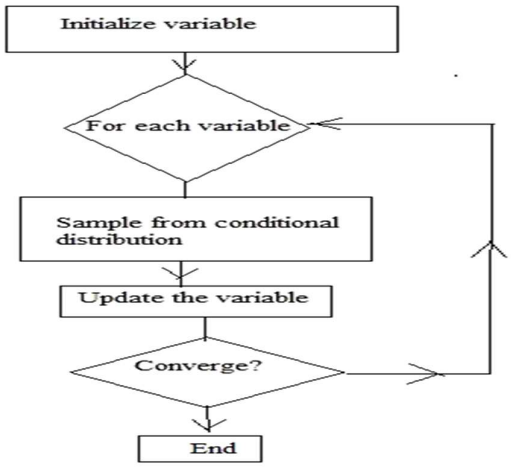

Figure 3 – Flowchart for Gibbs Sampling Approach

The following algorithm structures the procedures followed in the Gibbs Sampling approach

Algorithm-1: GIBBS Sampling Procedure
Input: Word vector of the form $X=(X_1, X_2, ..., X_n)$ and the Morfessor Model. Output: Morpheme length segmented tokens

```
Initialization:
    Periodic_Interval ← Rule initialized models (surface words)
    Boundary ← Evaluate (Periodic_Interval)
    Out_Reach ← Peridoc_Interval
    Segmented_Count ← find_Counts(Out_Reach)
    For i → Counts do
        For j → Length (morphemes) do
            For k → segments (Words) do
                M_x^- ← Process (Prob (segments[j]_k^-))
                M_y^+ ← Process (Prob (segments[j]_k^+))

                If NoBoundary Markers FoundAt(k) then
                    Check boundry with Out_Reach
                    Probability ← M_y^+ / (M_y^+ * M_y^+)
                Else
                    Place the boundary at k
                    Probability ← M_y^+ / (M_y^+ * M_y^+)
    UpdateOut_reach (Periodic_Interval)
    ElangateSegmentCounts (Surface Words)
```

Algorithm 1: GIBBS Sampling Procedure

This paper use Morfessor 2.0 system [16] (A.Rouhe et al, 2020) over the unigram segmented model (Unsup-uni), For both informal Tamil and Deutsch (German Language), the following procedures are carried out: (a). Boundary (b) Probability variation with variable P_0^A (c) Probability variation with lexeme through Morfessor system. During each iteration, an incremental gap in the morpheme boundary is introduced whenever the model tries to saturate for longer iterations.

During the procedure at stages (a), (b) and (c), an additional co-existence dataset known as Addon_data is attached. Addon_data is an unsegmented dataset for supporting the testing while performing the incremental training the systems. Though negligible amount of contribution is presented from the UnSup_Uni dataset unigrams, it is evident to cumulate the contributions from the Morfessor unigrams.

In the above algorithm, the Out_Reach corresponds to Boundary and the segmented length of the morpheme under consideration. For each morph length (x_i) accumulation in the boundary probability is studied at each marker position, resulting in (x_i / length) close to 0.76. For finding out - Prob (segments[j]$_k^-$), the following 2 steps are done: (i). iterating the Gibbs sampler [20] periodically (ii) non-parametric random change and iteration. Setting the morpheme length under the prefixed M_x^- every time the periodic distribution is updated. In our experiment the sampler is allowed to run for 100,000 (random) iterations. Samples collected after the each 1000 iterations are analyzed for the sign of any convergence possibility.

As provided by the Gibbs sampler, the convergence parameter (α), is kept on added recurrently. The base distribution and morpheme extraction probability are kept apart from the posterior and assertions for the classification. The grammar annotation context which needs to be provided for the Morfessor 2.0 is fed along with the Addon_data dataset. The expected length in the grammar is evaluated with respect to the following parameters namely Precision (P), Recall (R) and F-Score (F).

5.3. Data Analysis

The conjugate and annotated data corpus EMILLE (Xiao et al, 2014) [17] along with the Addon_Data corpus of voluminous data pertaining to Informal Tamil and German is considered. The original Unicode Transformation Format (UTF-8) and the transliteration is followed with precision, recall and

F-score and plotted. For training Addon_data a corpus segment of 35,000 inherent words are tested with the empirical reference corpus. Figure-4 shows the morph count distribution of both the informal Tamil and Deutsch (German Language).

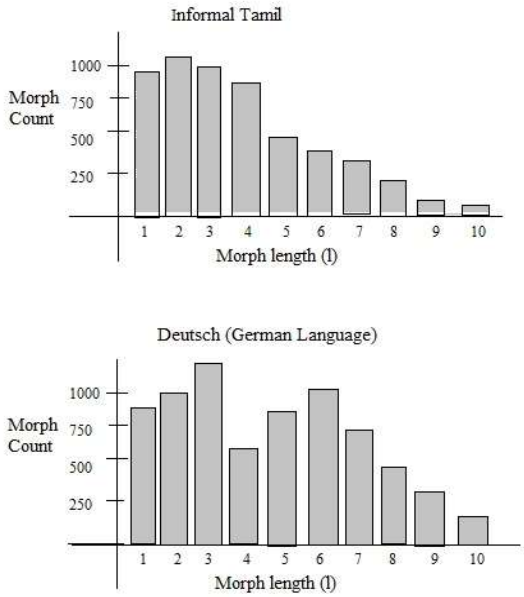

Figure 4 – Morph length and Morph count / Informal & Deutsch Language

6. Few Observations

For Tamil, most morphs have a length of 1 to 4 characters. The Boundary model and UnSup-Uni perform well within this morpheme range. For lengths of 1 to 4 characters, both models outperform Morfessor in terms of F-score. However, the performance of Boundary model and UnSup-Uni steadily decreases and starts to fall behind Morfessor for lengths greater than 5 characters. This behavior is somewhat expected, as UnSup-Uni models are contextually coherent to length priors and perform poorly if the assumed morph lengths deviate significantly from the actual assumed range. In contrast, Morfessor maintains consistent performance across the entire length range of 1 to 10 characters. This suggests that Morfessor' decline towards length priors, even with significant changes in the expected morph length. Overall, unsup-uni-p0-len achieved the best performance with an F-score [18] of 48.83% compared to the other models in this task.

Deutsch typically exhibits morph lengths ranging from 2 to 8. Morfessor outperforms both unsupervised unigram models (unsup-uni-p0-len and unsup-uni-p0-lex-len) across all length ranges except for lengths 1 and 2. When compared to Tamil models across various length ranges, unsupervised unigram [19] models generally show poorer performance. Overall, Morfessor achieves a higher F-score (43.63%) compared to unsupervised unigram models in this task. Morfessor also demonstrates superior performance, particularly in the more frequent morph length range of 5-8.

The result in Table-1 & Table-2, intimates that UnSup-Uni model with respect to the posterior probability and the inclination towards the parametric adaptation. For informal Tamil Language, the UnSup-Uni model fits well and the linearity in the morph length decline as the number of segmented support level is also drastically degenerates. For Deutsch (German language), the morph length parameter is partly linear dependant and partly degrade as the segmented morphemes vanish.

Table 1: UnSup-Uni model based values

"UnSup-Uni" model Boundary parameter based analysis						
Morph Length	Informal Tamil			Deutsch (German Language)		
	Precision (P)	Recall (R)	F-Score (F)	Precision (P)	Recall (R)	F-Score (F)
1	17.04	72.98	27.05	08.07	74.67	15.82
2	17.76	49.08	26.01	09.06	52.60	14.39
3	15.86	33.92	21.19	08.19	33.95	13.67
4	16.33	22.86	19.27	06.83	24.68	12.72
5	14.75	18.46	17.21	06.07	20.56	12.21
6	17.63	14.50	16.91	08.45	16.59	11.52
7	16.65	12.82	15.28	08.26	14.66	10.87
8	17.11	10.47	13.19	09.28	14.83	11.42
9	15.79	13.31	5.53	09.69	13.79	11.64
10	15.98	11.28	13.67	06.19	11.07	11.70

Table 2: Morfessor boundary based values

"Morfessor 2.0" Boundary parameter based analysis						
Morph Length	Informal Tamil			Deutsch (German Language)		
	Precision (P)	Recall (R)	F-Score (F)	Precision (P)	Recall (R)	F-Score (F)
1	49.45	42.57	45.98	29.33	71.36	42.83
2	47.32	41.81	42.87	28.59	70.84	41.05
3	47.61	40.70	44.93	30.48	70.57	42.83
4	49.10	40.42	45.91	30.73	70.71	43.56
5	51.24	41.46	45.82	30.88	71.30	43.86
6	50.70	40.84	44.98	30.58	72.96	42.88
7	49.39	41.53	45.63	32.35	70.92	46.43
8	49.12	40.48	45.03	32.34	70.57	41.52
9	49.24	40.40	45.46	30.87	71.69	42.95
10	47.39	39.39	44.78	30.45	70.84	41.47

Figure 5 and Figure 6, illustrates the chart as equivalent to the Table-1: "UnSup-Uni model based values" and Table-2: "Morfessor boundary based values" respectively.

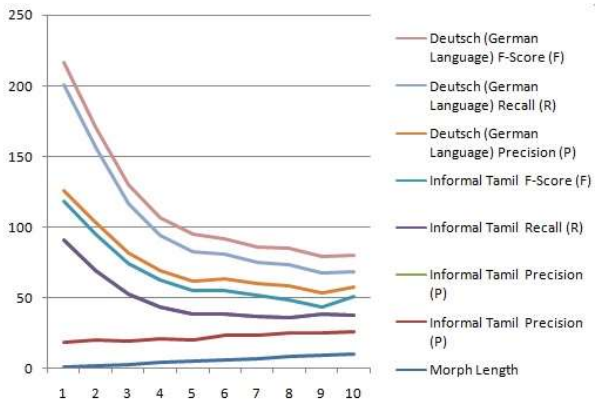

Figure 5 – Performance metrics of parameters based on UnSup-Uni model

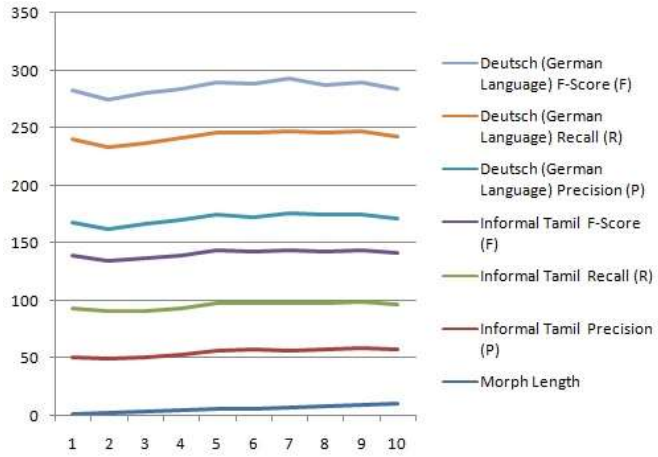

Figure 6 – Performance metrics of parameters based on Morfessor 2.0 model

7. Conclusion:

This article primarily investigates the knowledge of morph length affects the accuracy of morphological segmentation in agglutinative languages. To this end, a morphological segmentation based models on a Bayesian approach incorporating a posterior distribution over segmented morpheme and the length parameter. The results confirm that this knowledge supports the prediction accuracy manifold. For Informal Tamil Language, with an F-score of 48.83%, compared to Morfessor, it would be interesting to explore models and priors that address informal language dialectal changes. The Incorporation of Morfessor enhanced the addition of language-specific information. A comprehensive quantitative and qualitative evaluation using three metrics, Precision (P), recalls (R) and F-Score (F) for the two languages spanning the typological. The outcome shows the “Morfessor” and “Boundary parametric UnSup-Uni” both were able to support the analogy of co-existence of Dravidian languages in syntactical aspects.

Acknowledgement:

The authors thank the efforts and research contributions from various language researchers and the repositories that helped us supplying corpus during the test run.

References:

1. Eskander, R., Callejas, F., Nichols, E., Klavans, J. L., & Muresan, S. (2020, May). MorphAGram, evaluation and framework for unsupervised morphological segmentation. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 7112-7122).

2. Rouhe, Aku, Stig-Arne Grönroos, Sami Virpioja, Mathias Creutz, and Mikko Kurimo. "Morfessor-enriched features and multilingual training for canonical morphological segmentation." In Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 144-151. 2022.

3. Eskander, Ramy, Francesca Callejas, Elizabeth Nichols, Judith L. Klavans, and Smaranda Muresan. "MorphAGram, evaluation and framework for unsupervised morphological segmentation." In Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 7112-7122. 2020.

4. Mamouras, Konstantinos, and Agnishom Chattopadhyay. "Efficient matching of regular expressions with lookaround assertions." Proceedings of the ACM on Programming Languages 8, no. POPL (2024):

2761-2791.

5. Talukdar, D., and M. Tsubokura. "Numerical study of natural-convection from horizontal cylinder at eccentric positions with change in aspect ratio of a cooled square enclosure." Heat and Mass Transfer 58, no. 5 (2022): 849-871.

6. Batsuren, K., Bella, G., Arora, A., Martinović, V., Gorman, K., Žabokrtský, Z., Ganbold, A., Dohnalová, Š., Ševčíková, M., Pelegrinová, K. and Giunchiglia, F., 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. arXiv preprint arXiv:2206.07615.

7. S. Thivaharan., G. Srivatsun. and S. Sarathambekai., "A Survey on Python Libraries Used for Social Media Content Scraping," 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2020, pp. 361-366, doi: 10.1109/ICOSEC49089.2020.9215357..

8. Hahn, Michael, and Yang Xu. "Crosslinguistic word order variation reflects evolutionary pressures of dependency and information locality." Proceedings of the National Academy of Sciences 119, no. 24 (2022): e2122604119.

9. Lin, Yi-Tang. Statistics and the language of global health: Institutions and experts in China, Taiwan, and the world, 1917–1960. Cambridge University Press, 2022.

10. Feng, Steven Y., Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. "A survey of data augmentation approaches for NLP." arXiv preprint arXiv:2105.03075 (2021).

11. Sarveswaran, Kengatharaiyer, Gihan Dias, and Miriam Butt. "Thamizhi Morph: A morphological parser for the Tamil language." Machine Translation 35, no. 1 (2021): 37-70.

12. Manoharan, Geetha, Subhashini Durai, Gunaseelan Alex Rajesh, and Sunitha Purushottam Ashtikar. "A Study on the Application of Natural Language Processing Used in Business Analytics for Better Management Decisions: A Literature Review." Artificial Intelligence and Knowledge Processing (2023): 249-261.

13. Thivaharan.S, Srivatsun.G, "Maximizing the Prediction Accuracy in Tweet Sentiment Extraction using Tensor Flow based Deep Neural Networks", IRO Journal of Ubiquitous Computing and Communication Technologies (IROUCCT), June 2021, Vol.03, Issue.02,pp.61-79,ISSN: 2582-337X.

14. Klavan, Jane, and Ole Schützler. "The complexity principle and the morphosyntactic alternation between case affixes and postpositions in Estonian." Cognitive Linguistics 34, no. 2 (2023): 297-331.

15. Le, Xuan-Bach D., Corina Pasareanu, Rohan Padhye, David Lo, Willem Visser, and Koushik Sen. "Saffron: Adaptive grammar-based fuzzing for worst-case analysis." ACM SIGSOFT Software Engineering Notes 44, no. 4 (2021): 14-14.

16. Sharma, D. K., Bhopendra Singh, M. Raja, R. Regin, and S. Suman Rajest. "An Efficient Python Approach for Simulation of Poisson Distribution." In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 1, pp. 2011-2014. IEEE, 2021.

17. Zhou, Yu-Cheng, Zhe Zheng, Jia-Rui Lin, and Xin-Zheng Lu. "Integrating NLP and context-free grammar for complex rule interpretation towards automated compliance checking." Computers in Industry 142 (2022): 103746.

18. Yamaguchi, Atsuki, George Chrysostomou, Katerina Margatina, and Nikolaos Aletras. "Frustratingly simple pretraining alternatives to masked language modeling." arXiv preprint arXiv:2109.01819 (2021).

19. Srivatsun, G. and Thivaharan, S. 'Modelling a Machine Learning Based Multivariate Content Grading System for YouTube Tamil-post Analysis', Journal of Intelligent & Fuzzy Systems, vol.45, Issue: 6, pp. 11925-11936, DOI: 10.3233/JIFS-222504, 02 December 2023 (Print), IOS Press central Library, ISSN 1064-1246 (P), ISSN 1875-8967 (E).