Data Preprocessing

# Converting Tamil from a Syllabic to a logosyllabic script

* I used the labelled Tamil database provided by the [Institute of Formal and Applied Linguistics (UFAL)](file:///C:\Users\keetu\AppData\Roaming\Microsoft\Word\Institute%20of%20Formal%20and%20Applied%20Linguistics%20(UFAL)) available here: <https://ufal.mff.cuni.cz/~ramasamy/tamiltb/0.1/introduction.html#1.2.Data> [1]
* The dataset consists of Tamil news articles tagged by part of speech. Some of the morphemes including clitics and postpositions have been separated already. However, the dataset does not separate noun and verb morphemes. Note that Tamil word morphology is exclusive to nouns and verbs.
* A few grammar forms were also missing in this dataset such as adjectival participle verbs.
* Also Note that Modern Tamil has clearer rules for morphology than Old Tamil. There are also certain morphological differences for example
  + The presence of four plural suffixes instead of one
  + The presence of only two tenses (past and non-past) etc

But despite the differences, modern Tamil still demonstrates a morphological continuity from old Tamil, for example – both have similar noun cases. [2]

See the image below for the conll file provided by the database. Each word was tagged by their part of speech (POS)

Text

Description automatically generated

1. Tamil\_Morpheme\_List.ipynb -

I considered using unsupervised algorithms such as the Morfessor algorithm (combination of minimum description length and maximum a posterior algorithm) or a Nested Pitman-Yor based algorithm for morpheme segmentation to segment the noun and verb morphemes. However, the literature review revealed that using these algorithms for Tamil had an F score of around 50%. Reasons they failed included morphophonemic changes at morpheme boundaries and failing to identify character combinations that needed to be considered as a single character. [3][4] This is why I filtered the dataset for each noun and verb part of speech using regular expressions and manually created a morpheme list.

1. Convert\_clitics.ipynb

I identified other parts of speech that were already separated such as postpositions and quantifiers and added them to the clitics list. Then, I gave them each a unique numeric identifier and replaced them in the complete dataframe.

1. Convert\_lemmas.ipynb

I identified root words, gave them each a unique numeric identifier and replaced them in the complete dataframe. I converted root words prior to converting morphemes because I did not want to replace morphemes occurring in the root words.

1. Convert\_morphemes.ipynb

I gave unidentified morphemes as well as letters appearing in proper nouns identifiers. I assumed that proper nouns are spelled syllabically in a logosyllabic language.

1. Convert\_logosyllabic\_words\_to\_sentences.ipynb

I converted the logosyllabic words to sentences using punctuation label

Chart, diagram, schematic

Description automatically generated

[1] Ramasamy, L., & Abokrtsk, Z. (2011). Tamil dependency parsing: results using rule based and corpus based approaches [Dataset]. Springer-Verlag. <http://portal.acm.org/citation.cfm?id=1964799.1964808> License: <http://creativecommons.org/licenses/by-nc-sa/3.0/> <http://creativecommons.org/licenses/by-nc-sa/3.0/legalcode>

[2] Thomas Lehmann. 18 Dec 1997, Old Tamil from: The Dravidian Languages Routledge

[3] Kumar, A., Padro, L., & Oliver, A. (2015). Unsupervised learning of agglutinated morphology using nested Pitman-Yor process based morpheme induction algorithm. 2015 International Conference on Asian Language Processing (IALP). https://doi.org/10.1109/ialp.2015.7451528

[4]  Sheshasaayee, A., & Deepa, A. (2014). A structured analysis on analysis on morpheme segmentation for agglutinative languages Dr. Ananthi Sheshasaayee1 , Angela Deepa.V.R. International Journal of Advanced Technology in Engineering and Science, 02(01).