INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD
R&D SHOWCASE 2021
Technology, Social Impact

Building tools for psycholinguistic tasks in Indian languages

ABSTRACT

The initial aim is to find a way to generate Nonwords and syllabify (non)words in Indian Languages to be used in psycholinguistic tasks like building an Aphasia Battery.

Nonwords are sequences of sounds in a language which sound phonologically close to actual words in a language but do not have any meaning by themselves. It is difficult to exhaustively list out the phonotactics of a language to make a rule-based model for such a generation task, especially when the database needs to be big and varied enough for lots of psycholinguistic filters. One of the filters involves syllabification. It is the task of breaking down a (non)word into syllables, which are sub-word segments with a vowel and optionally some consonants.

Thus, the study tests the validity of a long short-term memory (LSTM) based recurrent neural network approach to generate nonwords and reviewing rule-based approaches used in the past for syllabification in Indian Languages.

OBJECTIVE

The constraints considered while finding a method to generate nonwords were:

- It should understand the phonology of the target language.
- It should be able to handle different Indian Languages without a lot of changes to the model.
- It should be filterable on psycholinguistic factors like (non)word neighbors, number of syllables per word etc.
- The syllabification model should automatically break down complex words like a native speaker would. Example: आत्म-कथा = ɑ: t̪ . m ə . k ə . t̪ʰ ɑ:

Reading Aloud Non-Words	Comprehension of oral spellngs	Spelling Aloud	Word & Non-word Judgement	Writing Practice: Copying Letters, Words, Nonwords	Writing to dictation
खूई	पदयाही	उटाना	फागना	रक्क	झदमी
पैंध	बुनदर	शूत	दूबन	आलो	सीयात
पूका	हुरकी	तरब	पुलम	आवड़	भुमन्य
गोर	मोवकबा	कोफ़	जूरना	उनजी	लूडरू
शोता	ख़ालती	सपेत	परसन	रई	गोनडन
कूना	पन्क्यान	कात	लुगख	लीडी	महाना
दोख	भज़क़ूली	बर	बछास	ठला	अशसी

Some generated nonwords, arranged by tasks in HRWIT (Hindi Reading Writing Indore Test)

METHOD

To syllabify, we mirrored previous research to break down simple words and plan to use a dictionary-based approach to tackle compound and complex words. While, to generate nonwords we used a multi-layer LSTM network, but instead of using it on a word level, it was refocused onto a sub-word level to understand the phonotactics of the target language. This was done by:

- using training data which was curated for regular phoneme representation, sub-word division etc.
- tuning hyperparameters for the optimum number of RNN layers, hidden units, and character-level embedding dimensionality. The best performing model was isolated using K-fold validation.

The model (pilot-tested by Hindi speakers) gave us 95% accuracy (measured by percent of tokens generated that follow Hindi phonology).
We next aim to run validation experiments on the nonword database generated by testing it on native speakers (aphasic and non-aphasic) apart from using the same pipeline for other Indian languages.

Authors: Mukund Choudhary, Jashn Arora, Dipti Misra Sharma, Bapi Raju Surampudi & Dr. Apoorva Pauranik, Indore

Research Centers: Cognitive Sciences Lab & LTRC