Formal Grammar Generating Procedure for Optimising Speech Recognition and Voice Control in Domain Specific Vocabulary Applications

Shashwat Goel, Delhi Public School RK Puram, New Delhi
Wrik Karmakar, South City International School, Kolkata

We develop a novel approach for speech recognition and control with increased robustness, tolerance to failure and applicability to critical domains like aviation, security, space, and healthcare where training regimes exist. Furthermore, recognition systems independent of English grammar are immensely useful for non-native speakers, who then just need to memorize a smaller vocabulary and its meaning. We propose an automated method of generating concise sets of user-friendly, unambiguous task-oriented commands from text corpora, to improve discernibility.

We pre-process varying datasets to produce uniform input. Customized Part-of-Speech tags are used to suit whatever industry required. Lexicons for each tag are processed separately hereafter. Word embeddings provide a numerical visualization of this semantic interpretation of words. Thus, each lexicon separated for different commands is plotted in a 300-dimensional vector space by training embeddings using a GLoVe system enriched specifically by troponym relations, a feature of verb pairs that most accurately describes verb relations. K-means clustering is performed to group words with similar meaning. The 'K' to run K-means clustering is chosen by the elbow method, that is, a F-Test (Number of clusters vs Percentage variance) graph.

Each cluster is represented by a central word, and the clustering condenses the vocabulary. To achieve this we consider 5 parameters - vocabulary size, phonetic distance, loss of meaning, familiarity, and command complexity. Among these the most innovative was avoiding phonetic collisions between selected words. Phonetic similarity was evaluated by ABX testing, traditionally used to evaluate compression quality.

For this, a congruence to the NP-Hard Maximum-Independent Set problem was proven by taking words as nodes and augmenting edges between phonetically or semantically similar pairs. To convert from a node-weighted graph to an unweighted one, we proved that if a particular node is selected, all its copies can be selected for a better solution. Consequently we also proved that the weight is accounted for in this process. Observing the sparsity of the graph, a SAT Solver is used. Rapid Automatic Keyword Extraction (RAKE) was used on WordNet definitions to represent the remaining words in the cluster, ensuring minimal loss of semantic meaning.

The vocabulary and syntax are finally to be parsed into a grammar-assisted speech recognition model. We propose an innovative method for evaluation. The dataset and its formalized version are converted from text-to-speech with noise added from the AURORA database. This speech is then converted back to text, and accuracy compared.

We have currently applied the procedure on the Cornell Tennis Commentaries dataset. A 53% reduction in the Noun lexicon, and 37% reduction in the Verb lexicon was achieved. Given that the dataset is not a command/control dataset, the results will only improve once our request to access such private datasets is approved. We hope to create a custom speech recognition using the CMU Sphinx software to fully test our work in the future

Giving a computational basis to what has previously been attempted manually to create callsigns and radiotelephony procedure, we empower the revolution of intelligent systems that our future shall embrace.