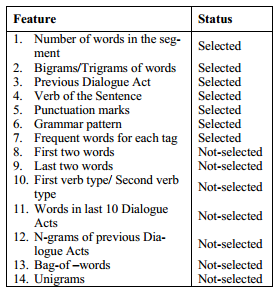
**Feature Selection**

Features are the input given for the classifier as a vector for each word in the utterance. Classifiers take decision according to the defined set of features. Features can be extracted from the word itself, timing and the prosodic information .Features can be broadly categorized as time related, prosodic, word related and online features. So, selecting a correct set of features is inevitable in DA recognition. Our objective is to test the performance of features already identified in related work for English and explore some new features which are unique to Sinhala. Fortunately, we have identified several new features exclusive for Sinhala.

We have identified 14 heavily used features from related work and selected only seven of them by considering the applicability to Sinhala and few other concerns like duplicability, unavailability, erroneous and ineffectiveness. The following table illustrates this scenario well.



Since, we are using the Bigrams as a feature, features such as first two words and last two words can be omitted. First verb type/ second verb type is omitted due to the unavailability of Sinhala PoS tagger. Taking previous Dialogue Acts as features can introduce a cumulative error. So, previous Dialogue Act has not been taken as a feature. Unigrams are omitted due to its ineffectiveness for long utterance and we have taken the bag of words as the base case.

**Exclusive features of Sinhala**

Exclusive features of Sinhala can be good indicator for the classifier to do the classification process. The following are some identified unique features of Sinhala.

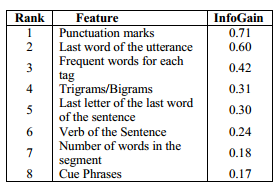
* **Last letter of the last word of the utterance**- In Sinhala , Yes/No questions ends with the letter ‘ද’(da), most of Request/Command/Order ends with one of the letters ‘න්’(n), ‘න’(na), or ‘ නු’(nu), most of Open questions end with ‘නේ’(ne).
* **Last word of an utterance –** mostly represent the verb of an utterance
* **Sinhala specific cue phrases -**

E.X: ඇත්තෙන්ම, **සහ, එසේම, වගේ, ඉතින්, නමුත්, එනිසා**

**Feature Set Selection**

Now, we need to identify actively contributing features and most effective combinations of features for the classification. From the aforementioned 10 features (7 general features and 3 Sinhala specific features) 8 were selected based on the performance evaluation. Because, it is computationally expensive to go through all the possible combinations of ten features compared to eight features.

We used WEKA Java library for classification task. To achieve above described task we used InfoGain Attribute Evaluator of WEKA and obtained the InfoGain values. Table 9 displays the results. The InfoGain value evaluates the worth of a feature by measuring the information gain resulted only by that particular feature. For example, a feature with an InfoGain value of 1 means that all of the information available in that feature contributes to classification, though it does not mean that the use of that feature alone is able to conduct the entire classification. The following diagram shows the InfoGain values in descending order.



By looking at the infoGain value, we can clearly observe that the Exclusive Sinhala features that we introduced also significantly contributing to the classification task.

**Best Performing Six Features**

* Punctuation marks
* Last word of the utterance
* Trigrams/Bigrams
* Last letter of the last word of the sentence
* Frequent words for each tag
* Cue Phrases