## Introduction

Intelligent agents have been utilized for many domains within the last decade. One of the crucial requirements of such intelligent agents (live chat based customer service providers, intelligent voice assistant agents etc.) is their ability to understand spontaneous dialogues.

To understand a spontaneous dialogue, it is important to model and automatically identify the structure of that dialogue, because it will make it easier to get a better interpretation of that spontaneous dialogue. How to model a spontaneous dialogue precisely is still an open issue, though some of the specific characteristics for modeling a spontaneous dialogue have already been identified. Among these clearly identified characteristics, “Dialogue Acts” hold an important place.

The process of identifying the Dialogue Acts (DAs) for a particular language consists of fixed set of steps [1]. That process is independent from the natural language used for the Dialogue Act Recognition. First and foremost step of the dialogue act recognition procedure is to identify the set of DA tags that is relevant for the task. After that, relevant informative features have to be computed from the speech signal. That is a very critical step since the accuracy of identifying the Dialogue Acts heavily depend on the identified feature set. And then DA models will be trained on these identified features set. To make the process of dialogue act recognition easier, the segmentation of the dialogues into utterances needs to be carried out independently, or alternatively realized during the recognition step with joint DA recognition and segmentation models.

Before going into the deeper level of information related to dialogue acts we first provide a brief introduction to the topic and cover some essential fundamental concepts related to dialogue acts such as illocutionary forces and speech acts to provide the background to the topic. The next section discusses existing corpora used for dialog act recognition for English. Different approaches for building a Sinhala corpus for dialogue act recognition are also discussed. The next section focuses on the dialogue act tag sets used in the process of dialogue act recognition on some of the standard corpora discussed before. It is important because almost all the research work done in the area of DA recognition considered those tag sets as a standard and used an appropriate subset of those DA tags for the research. The next section will briefly discuss the idea of inter-annotator agreement which is a useful statistic to measure the accuracy of tagging the utterances of the corpora. Following that, we will discuss the important topic of feature selection. The final section of this literature review focuses on the major standard classification techniques used for dialogue act recognition and the techniques used to rate the performance of classifiers.

## Speech Acts and Illocutionary Forces

A speech act in [linguistics](http://en.wikipedia.org/wiki/Linguistics) is an utterance that has performative function in language and communication [2]. In general, speech acts are acts of communication such as statements, requests, questions, apologies and thanking. These acts of communication are for expressing a certain attitude, and the type of speech act being performed corresponds to the type of attitude or intention being expressed. For example, a statement expresses a belief, a request expresses a desire, and an apology expresses a regret. As an act of communication, a speech act succeeds if the audience identifies, in accordance with the speaker's intention, the attitude being expressed. So dialogue acts are specialized versions of these speech acts. For example “Question” is a speech act, but “Yes-No-Question” is a dialogue act. Therefore although the number of speech acts is somewhat stable, usually ten, the number of dialogue acts depends. For example if the requirement is to process a questionnaire system, it is required to have different kinds of questions like yes-no-questions, open questions etc. However having different kinds of greetings is useless for that application. That explains how the set of dialogue acts and the size of the set depends on the application.

Austin [3] defines a dialogue act as the meaning of an utterance at the level of illocutionary force. The illocutionary force of an utterance is the speaker's intention in producing that utterance. Instance of a culturally defined speech act type is known as an illocutionary act, it is characterised by a particular illocutionary force. It has several types of acts, such as Asserting, Promising, Excommunicating, Exclaiming in pain, Inquiring and Ordering. For example, if we consider a speaker who asks “How is that work going on?, Is it finished yet?” as a way of enquiring about the work, his or her *intent* may be in fact to make the person to finish the work. Thus the illocutionary force of the utterance is not an inquiry about the progress of the work going on, but a force for the work to be finished. A different way to define what is a dialogue act is, “a dialogue act is a specialized speech act”.

### Identifying the Tag Set

As discussed in the above sections dialogue acts are the basic building blocks for the process of spoken language understanding in human conversations. So selecting an appropriate dialogue act tag set is the crucial first step in processing conversational speech. This heavily depends on the language, culture and the context of the target application/task. Kral [1] identifies the following three requirements for selection of tag set which can be applied to any language and the any context.

1. The DA tags should be generic enough to be useful for different tasks, or at least robust to the unpredictable variability and evolution of the target application.
2. The DA tags must be specific enough to encode detailed and exploitable characteristics of the target task.
3. The DA tags must be clear and easily separable, in order to maximize the agreement between human labelers.

Although as the above three rules explain, the selection of a tag set heavily depends on the context of the target application/task, there are some tag sets that are usually used as common base lines for most tasks. In a study, what usually happens is, first these common tag sets are studied and then target specific DA tag sets specific for the context are derived.

The natural language processing community with the guidance of the Discourse Resource Initiative [4] has designed the Dialogue Act Markup in Several Layers (DAMSL) tag set. With this tag set, they have targeted to provide a domain-independent universal framework for dialogue annotation. Its annotation scheme was a composition of four statistically independent (orthogonal) dimensions as follows.

1. Communicative status (defines whether the utterance is interpretable, abandoned or a self-talk.)
2. Information level (provides an abstract characterization of the content of the utterance.)
3. Forward looking functions (provides a classification in a way as actions in Searle’s [2] speech act theory.)
4. Backward looking functions (defines the relationship between the current utterance and the previous dialogue acts.)

So it is possible to combine tags from different dimensions to create new tags appropriately. After DAMSL, most of the related work used an adaptation of this tag set. The Switchboard DAMSL (SWBD-DAMSL) tag set [5] is one of the widely used adaptation. It was initially designed for the domain of telephone conversation classification.

Adapting the DA tag classes (approximately 60 tags in orthogonal dimensions) of DAMSL, SWBD-DAMSL has derived 220 different dialogue act tags to tag 205,000 utterances of the Switchboard corpus. But after clustering the rarely occurring tags (130 of the tags occurred less than 10 times) the final tag set consisted of 42 classes of DA tags.

The Meeting Recorder Dialogue Act tag set [6] is another commonly adapted tag set. It was an adaptation of the SWBD-DAMSL tag set, to classify the utterances of the ICSI (International Computer Science Institute) meeting corpus [7] that consists 72 hours of naturally occurring multi-party meetings manually tagged with DAs and adjacency pairs. The importance of this tag set comes through the characteristics of the corpus because it contains natural meetings that contain regions of high speaker overlap, affective variation, complicated interaction structures, abandoned or interrupted utterances and other interesting turn-taking and discourse-level phenomena. So this tag set can be considered as a more generalized tag set compared to the other tag sets derived upon corpora that were built on top of data collected using artificial scenarios.

Another popular dialogue act tag set is the VERMOBIL DA tag set [8]. It consists of 42 dialogue acts and 18 dialogue acts in the illocutionary level. While all the above mentioned DA tag sets are developed focussing on the English this tag set was initially designed to fulfil the requirements for the German language. So it was initially designed to facilitate German and English languages. For the determination of dialogue acts of English utterances they have used the keywords while for German utterances they have used the micro and macro structural information. The basic idea of the system is to homogeneously model preference rules by taking the information from various sources. Using that decision tree, it is possible to clarify the relationship between dialogue acts, and during the tagging process the tree is parsed from root to the leaves.

For almost all the research purposes, people have used combinations and subsets of the above mentioned tag sets. To determine the size and the composition of the appropriate dialogue act tag set, it is important to consider the size of the corpus, the context of the task etc. The natural language that was used for the study of dialogue act recognition is another important factor when selecting the tag set. Because other than the VERMOBIL tag set all the other commonly used standard tag sets are defined considering the characteristics of the English language. So it is really important to consider the language native characteristics before defining the DA tag set for the study.

If we consider the Sinhala language, although there are very few similarities between Sinhala and English we can consider it as a different language with a different set of language characteristics. For example, in English it is an easy task to categorize the utterances into the categories, commands, orders and requests, because there is significant difference between command utterances, order utterances and request utterances. For example, in a request utterance the word “please” is commonly used while in an order it is rarely used. But with the native language characteristics of Sinhala it is hard task to find the separation between these three categories. With the prosodic information it might be possible, but only with the lexical and syntactical information of the utterances, task is almost impossible. So the best thing to do is combine these three categories into a single dialogue act and use it.

### Existing Corpora

Today DA modeling techniques are widely used in the speech translation systems. That is, human-to-human communication through a machine conducting language translation. Spontaneous dialogue speech corpora are essentially important to model relevant features of spontaneous speech, such as pauses, hesitations, turn-taking behaviors, etc and dialogue structures. There are several key corpora that have been used in most work on DA modelling. Among those few are publicly available for further studies while others are restricted. [See figure. 1]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Corpus** | **Utterance Count** | **Word Count** | **Distinct Words** | **Dialogue type** |
| SWITCHBOARD[9] | 223 606 | 1 431 725 | 21 715 | Conversational |
| VERBMOBIL[10] | 3 117 | 24 980 | 959 | Task-oriented |
| ICSI MEETING RECORDER[6] |  |  |  | Conversational |
| MAPTASK[11] | 26 621 | 152 705 | 2 502 | Task-oriented |

### Figure. 1: Available Corpora

**VERBMOBIL**

It is a German research project, project that aims at translating the spontaneously spoken dialogues robustly and bi-directionally for German/English and German/Japanese. It produced a corpus of 168 English annotated task-oriented dialogues. In order to tag the VERBMOBIL corpus they used a total of 46 tags, which are then further clustered into 26 top-level tags.

**SWITCHBOARD**

This corpus comprises 1155 annotated telephone conversations that have greater variability of topics. Due to that, SWITCHBOARD corpus exhibits greater semantic variability than any other corpus created that time. Therefore it has been a more difficult problem for accurate DA modeling. This corpus was initially tagged with 220 tags. 130 of those tags that occurred less than 10 times have been clustered. Finally that lead to the 42 larger tag classes.

**ICSI MEETING RECORDER**

This project had a corpus of over 180,000 hand annotated dialog act tags and accompanying adjacency pair annotations for roughly 72 hours of speech from 75 naturally-occurring meetings. For tagging the Meeting Recorder corpus used 65 tags. Most of tags they adopted from the SWBD-DAMSL since it fits with the corpus.

**MAPTASK**

Design of this corpus mainly focused on allowing the investigation of a range of issues relevant to both psychological models of human language production and comprehension and speech technology, especially as the focus on effort switched to more natural, unconstrained speech. MAPTASK corpus runs about 18 hours of speech, which have generated 152 705 word counts. Word lists containing all the feature names were also elicited from all speakers, along with a number of 'accent diagnosis' utterances. That contains 128 task-oriented dialogues and it uses 12 distinct DA tags.

**Corpora for Sinhala**

“Sinhala” is the native language for 80% of Sri Lankan citizens. In order to perform linguistic research on Sinhala language, the main pitfall is the lack of a standardized corpora for Sinhala. To create a corpora for Sinhala language one can use several techniques to collect data such as,

* Create a Sinhala chat tool and collect chat data
* Create a tool to extract data from Sinhala News Paper / Novels etc.
* Create a tool to extract Sinhala Subtitles for foreign movies/ TV series.
* Collect telephone conversation data

Among above points, the last point is a very difficult task to perform, because it might require lot of tools to capture a high quality data set. As well as speech-to-text conversion also needed. But for sinhala no robust method is there to perform such task. But the initial 3 points look a bit easier compared to the last technique. Creating a chat tool and collecting those data would be bit time consuming. But other remaining two facts are comparably easy to other methodologies. But finding Novels and Newspapers will be bit difficult compared to finding subtitles files on the internet, because there are huge archives existing for Sinhala subtitles for foreign movies and TV shows. In order to get those files from such archives one can create a small web crawler and then create a small program in any programming language to extract and store those sentences in a database easily.

### Feature Selection

Akker and Schulz [12] identify features as the input given for the classifier, as a vector for each word in the utterance. Features can be extracted from the word itself, timing and the prosodic information. They have further identified 4 major categories of features as they present in an utterance.

1. Time Related Features
2. Word Related Features
3. Prosodic Features
4. Online Features

Time related features are derived from the start and end times of the utterances in the corpus. ***Pause between two words*** is one feature. It is the difference between the start time of the current word and the end time of the last word by the same speaker. ***Duration of the word***, ***Mean duration of the word*** and ***Relative duration*** of the word are other word duration oriented features which are extracted from the corpus. For deriving these features from the corpus, the utterance itself needs to have information about the speaker, the duration of each word, etc.

Words can act as features themselves. These are categorized under the word related features. ***Current word***, ***next word*** and ***previous word*** are 3 most commonly used features. Since most classifiers cannot deal with strings directly, most of the time the feature of the string is converted into a nominal feature for each word. Akker and Schulz [12] have only used words in the corpus that occurred more than 100 times as nominal feature. Using the same procedure next word and previous word features can be derived from the same utterance.

Stegeman and Akker [13] describe different perspectives of the above 3 features, considering the part-of-Speech (PoS) that the current, next and previous words are located in. PoS (Part-of -Speech) is a linguistic category of words which is generally defined by thesyntactic ormorphological behaviour of the lexical item in question. Commonly *Verbs, Nouns, Adjectives, Adverbs WH-Questions* are used as this categories. Feature ***PoS Current word*** is derived by a tagger from the current word and surrounding 6 words. The tagger used Penn Treebank English tag set containing 37 tags [14]. ***PoS previous word*** feature derived from the words preceding the current word while ***PoS next word*** feature from the words that are following the current word.

Since the 37 sized tag sets used for PoS features are too fine grained for DA tagging purpose, the Penn Treebank tag set is mapped to a 6 tag set : *Verbs, Nouns, Adjectives, Adverbs WH-Questions and Other.* Using above 6 tags, they have redefined the PoS features as ***PoS Reduced Current word, PoS Reduced Next word*** and ***PoS Reduced Previous word***. Some words in an utterance have more impact on the DA than other words in the same utterance. So as an extension to this reduction, ***PoS with Keywords*** have been introduced where some words in the corpus get their own tag.

Other than the aforementioned features, two other word related features can be identified as ***Word Repeat*** and ***Word Repeat 2,*** where the former is evaluated to *true* if the next word is same as the current word and the latter is evaluated to *true* if the previous word is same as the current word.

Prosodic Features are derived from the prosodic information stored in the utterances [15]. Word pitch and Energy information can be used as features after amplitude values of the signal have been normalized to the microphones [16]. ***Pitch Features*** and ***Energy Features*** can be evaluated using the *minimum, maximum* and the *mean* values of pitch and energy. ***Speech-flow Past, Speech-flow Future*** and ***Speech-flow Change*** featuresdefine the talking speed with respect to the other words in the surrounding phrases.

Apart from above features, Akker and Schulz [12] discussed 4 other features related to segmentation of the utterances. ***Number of words in previous segment*** feature is self-explanatory. ***Distance to the last segment*** feature is the number of words from the end point of the last segment. ***Relative position of word inside the segment*** and ***Time interval of current word to last segment*** are other segmentation related features.

In the set of experiments done by Akker and Schulz [12], they have identified the best performing feature set including *Pause between words, Mean duration of words, Specific current words, Previous words, Part of Speech information, Minimum and Mean of Energy, Speed-flow change* and *Length of a segment.*

Apart from above discussed commonly used feature sets, other researchers have done several experiments with different types of features as well. Rosset and Lamel [17] used a feature-vector consisting of *Speaker Identity, Number of utterances* and *First two words.* Lendvai [18] opted for not using DA tag of previous utterance as a feature for the current utterance, as it could introduce a cumulative error. *Utterance type, Presence/absence of Wh-Question* and *Subject type* were used as features by *Andernach* [19]. He also used two interesting features *1st verb type* and *2nd verb type* because of their potential of informativity on kinds of agents and actions. Similar use of above mentioned two features can be seen in [15] as *grammar patterns.*

**Classification Techniques**

Dialog acts (DAs) represent the functional building blocks of conversations and the classification of dialog acts corresponding to assigning DA types to the individual utterances[20].How these different DA types are defined depends on many factors such as goal of the application, size of the corpus and the experimental setup. Dialogue act classification is a special case of text classification where the text to be classified is the user utterance. The state-of-art in dialogue act classification is to use all available information sources from multiple perspectives [21], including:

1. Linguistic information that can be derived from the surface form of an utterance: lexical and collocational information.

Linguistic DA classification is based on the observation that different DAs use distinctive word strings. It is known that certain cue words and phrases [22] can serve as explicit indicators of discourse structure. Similarly, we find distinctive correlations between certain phrases and DA types. For example, 92.4% of the uh-huh’s occur in BACKCHANNELS, and 88.4% of the trigrams “<start>do you” occur in YES-NO-QUESTIONS.

1. perceptual information from multiple channels available to dialogue participants, including acoustic and prosodic properties of utterances as well as information from visual and other modalities

Prosodic and acoustic DA classification is based on number of factors such as utterance duration, pitch, pauses, energy, speaking rate and gender that are computed automatically from the speech signal. Prosodic information is vital for DA classification, because word-based classification suffers from recognition errors and some utterances are inherently ambiguous based on words alone. For example, some Yes-No-Questions have word sequences identical to those of statements, but can often be distinguished by their prosodic information [23].

1. contextual information obtained from the preceding dialogue context and dialogue structure, as well as global context properties like dialogue setting, knowledge about dialogue participants, and domain knowledge

Structural information like location of an utterance and the context of the utterance can be a strong predictor of the dialogue act.Based on the nature of the application, it is observed that the utterance position in a sentence as well as in a turn plays an important role when identifying its dialogue act. For example, an utterance such as “Hello” will occur at the beginning of a dialogue while an utterance such as “Have a nice day” will typically appear at the end. So, the position of utterances in a turn can also help identify the dialogue act; i.e. when there are several utterances in a turn, utterances are related to each other, and thus examining the previous utterances in the same turn can help correctly predict the target utterance. For example, the greeting “Welcome” and question “How may I help you?” could occur in the same turn.

A wide variety of machine-learning techniques have been used for DA classification tasks with various instantiations of feature-sets and target class encodings, and for dialogue processing, it is still an open issue which techniques are the most suitable for which task [25].For example, techniques based on n-gram language modelling were applied by Reithinger [25] to the Verbmobil corpus, with a reported tagging accuracy of 74.7%. Hidden Markov Models (HMM) have been tried for dialogue act classification in the SWITCHBOARD corpus by Stolcke [23], achieving a tagging accuracy of 71% on word transcripts. Another approach that has been applied to dialogue act recognition, by Samuel [26], uses transformation-based learning. They achieved an average tagging accuracy of 75.12% for the Verbmobil corpus. Keizer [27] used Bayesian Networks, applying a slightly modified version of DAMSL with an accuracy of 88% for backward-looking functions and 73% for forward-looking functions in the SCHISMA corpus. Lendvai [28] adopted a memory-based approach, based on the k-nearest-neighbour algorithm, and reported a tagging accuracy of 73.8% for the OVIS data.

The following are some of the popular classification techniques used in the literature.

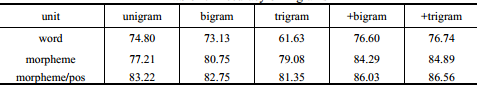
**N-Gram Model**

One of the simplest techniques used in DA classification is the n-gram. N-gram is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application- Wiki. N-gram models can be imagined as placing a small window over a sentence or a text, in which only n words are visible at the same time. The simplest n-gram model is therefore a so-called unigram model. This is a model in which we only look at one word at a time. The sentence “This is our literature review”, for instance, contains five unigrams: “This”, “is”, “our”, “literature”, and “review”. Of course, this is not very informative, as these are just the words that form the sentence. In fact, N-grams start to become interesting when n is two (a bigram) or greater. In similar fashion, a bigram can be thought of as a window that shows two words at a time such as {(This, is), (is, our), (our literature), (literature, review)}.

In the context of DA classification, N-Gram model is based on the assumption that the current dialogue act is explicitly determined by k preceding dialogue acts [29]. Therefore the candidate for the n-th dialogue act is chosen by the principle

Capture.PNG

Here the conditional probabilities are extracted from tagged corpora by counting all existing DA sequences [34]. These are simply the number of occurrences of the sequence (cn−k+1, . . . , cn) in the training corpus, divided by the number of occurrences of a shorter sequence, (cn−k+1, . . . ,cn−1). The typical values for k would be 2(bigram) or 3(trigram). Using larger k values only make sense when longer dependencies are known to exist. The unigram models lack the disambiguation information while the trigram models seriously suffer from the data sparseness problem. As represented in the below table [30], the combination models such as the +bigram models and the +trigram models achieve much better performance than other models; because the combination models can alleviate the sparse data problem by using the unigram features, and utilize the enriched features by using the bigram features or the trigram features.



**Hidden Markov Models**

Hidden Markov Models (HMMs) have remained as one of the dominant statistical modeling techniques used in modern Dialog Act recognition systems. HMM is a tool for representing probability over sequences of observations where a process is modelled as two parallel sequences of states, out of which one is observable and the other one is hidden. In this model, the state is not directly visible, but the output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM give some information about the sequence of states. The Hidden Markov Model is based on following the two important assumptions related to the nature of the dialog Act [31].

* Limited Horizon Assumption - The conditional probability distribution of future states of a process depends only upon the present state, not on the sequence of events that preceded it.

markov1.PNG

* Time Invariant Assumption - Transition and generation probabilities remain stationary throughout the process.

first part.PNGtime.PNG

Here X = (X1 ,..., XT) is a sequence of random variables taking values in some finite set S = {s1, . . . , sN),the state space.

There are two main approaches of HMM widely used in practise known as Forward-backward algorithm and Viterbi algorithms[32].Forward-Backward is used if one wants to predict the most likely token at a given time. It takes every possible sequence into account and averages over them to find the most likely token at that time. Therefore the returned sequence in not a true sequence, but a collection of the most probable tokens considering all of the possible sequences. Viterbi is used to find the most likely sequence of events. This algorithms looks at every sequence and simply selects the sequence that is most likely to be happen.

**Naive Bayes Classifier**

The Naive Bayes Classifier is a simple probabilistic classifier based on the Bayesian theorem with the assumption of a high degree of independence between features (strong independence assumption). The naive Bayes classifier assumes that the value of a particular feature is unrelated to the presence or absence of any other feature, given the class variable (DA tag). For example, a dialog act may be considered a back channel question if it is a backchannel and a question. This classifier considers each of these features to contribute independently to the probability that this DA is a backchannel question, regardless of the presence or absence of the other features.From Bayes theorem, we have

bayes.PNG

Here fi belongs to features and C belongs to the classes (DA tag). So the most probable class can be written as

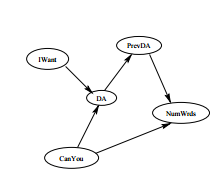
bayes1.PNG

According to the strong independence assumption, the probability of observing each feature for a given class is just the product of the probabilities of the individual features given in the class. So, the general Naive Bayes classifier can be written as [31]

naives.PNG

**Bayesian Network**

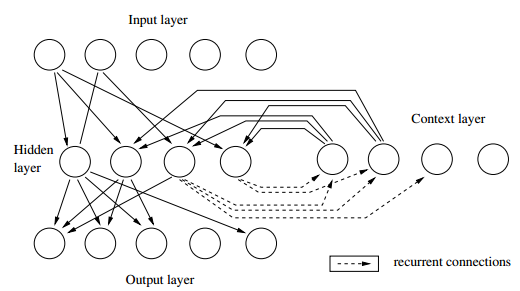
Bayesian networks or belief networks are an advancement over the naive Bayes classifier. Instead of assuming the features to be independent and the output to depend on the features, in this model the dependencies are specified by a directed acyclic graph (the dependency network). The output is regarded as an unknown feature that can both depend on and influence other features. This probabilistic graphical model is used to represent knowledge about an uncertain domain that encodes probabilistic relationships among variables of interest. Each node in the graph represents a random variable, while edges between nodes represent the probabilistic dependencies among the corresponding random variables. In a nutshell, the Bayesian probability of an event is the degree of belief in that event. The following diagram clearly illustrates the above mentioned scenario.



.

**Artificial Neural Networks**

Artificial Neural Network is one of the most frequently used neural network techniques, in general as well as in DA recognition. An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the multilayer structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. The input propagates through the network layer-by-layer. The following figure demonstrates this novel architecture of a simple neural network with one hidden layer and one context layer [32].



An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons [37]. Here it is called input weights and it is adjusted dynamically according to a learning rate until it reaches close enough to the target output. Simply, when a neural network is initially presented with a pattern, it makes a random 'guess' as to what it might be. It then sees how far its answer was from the actual one and makes an appropriate adjustment to its connection weights. ANNs, like people, learn by example. It cannot be programmed to perform a specific task. The training examples must be selected carefully otherwise useful time is wasted or even worse, the network might function incorrectly. The disadvantage is that because the network finds out how to solve the problem by itself, its operation can be unpredictable.

**Classifier Selection**

In order to select a suitable classifier to perform the classification, we need to get some measurements about how different classifiers perform on a set of utterances. A set of classifiers have to be trained and tested on a set of utterances. But again the test results affected by the feature set selection, number of training instances and classifier parameters. So it is convenient to use a fixed set of utterances as training set with selected fixed feature sets and without any classifier parameters.

**Precision and Recall**

In pattern recognition and data mining with classification, there are two significant parameters to evaluate the accuracy of the classification. The **Precision (**positive predictive value**)** and **Recall (**sensitivity**).** Precision is defined as the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved [33]. Precision can be seen as a measure of exactness or quality whereas recall is a measure of completeness or quantity. From the statistical viewpoint, when the scenario *all and only the relevant items are retrieved* istaken as the null hypothesis, absence of type I and type II errors corresponds respectively to maximum precision (no false positive) and maximum recall (no false negative).

*Definitions*

**F-measure**

F-measure combines the two measurements precision and recall and is taken as the harmonic mean of precision and recall [33]. F-measure reaches its best value at 1 and the worst score is 0.

The F-measure has been widely used in natural language processing literature. We can use F-measure to rank the classifiers against above mentioned feature sets and parameters.

### Inter-annotator agreement

After selecting the appropriate corpus and a tag set, the next critical step of the study is to tag the utterances of the corpus using the selected DA tag set.The tagging process should be done manually using real, non-biased people and for that, it is possible to use a single person or several people. It is really important to tag the corpus accurately in order to achieve best possible results in the final classifications. So to improve the accuracy of tagging, it is better to use several people for the task instead of a single person. So the better procedure is for several people to separately tag same part of the corpus manually and then review the results. It is possible to find an utterance where everyone used the same DA tag for an utterance and it is also possible to find the utterance is tagged using different tags by different users due to the ambiguities of the utterance or DA tag set.

So after tagging the corpus it is essential to find the comparison between the resultant tagged corpora of separate users. To evaluate this inter-annotator agreement value, Cohen [34] defined a statistic (Kappa value). The kappa value gives a percentage value of agreement between two annotators. But this method cannot be used to calculate the inter-annotator agreement between more than two users.To overcome that limitation Fleiss [35] defined a new statistic (Fleiss Kappa). That method is capable of finding the inter-annotator agreement between any number of annotators.

### Conclusion

In this review we have presented an extensive survey of Dialogue Act Recognition by data mining approaches and a brief review of related studies, research and techniques. We have explained the importance of recognizing Dialogue Acts in a conversation and the major steps involved with the process of identification. Extensive description about corpora used in various studies for different languages including English, German and French. We have compared the corpora available in the context against utterance count, word count and dialogue type. In this literature we pointed out that there are no corpus available for Sinhala language and we provided the approaches can be used to build a Sinala corpus.

Selecting features for the classifiers is a very important step in the process of dialogue act recognition using classifiers. We have summarized the feature selection criteria used by the studies done in the context. Then we discussed about different classification techniques can be used and we end the review with a discussion on selecting a classifier and the parameters we can use for measure classifier performance.

References

[1] Král, P., & Cerisara, C. (2012). Dialogue act recognition approaches.*Computing and Informatics*, *29*(2), 227-250.

[2] Searle, John R. *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press, 1985.

[3] Austin, John Langshaw. *How to do things with words*. Vol. 1955. Oxford university press, 1975.

[4] Allen, J., & Core, M. (1997). Draft of DAMSL: Dialog act markup in several layers. *Unpublished manuscript*, *2*.

[5] Jurafsky, D., Shriberg, E., & Biasca, D. (1997). Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*, 97-102.

[6] Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., & Carvey, H. (2004). *The ICSI meeting recorder dialog act (MRDA) corpus*. INTERNATIONAL COMPUTER SCIENCE INST BERKELEY CA.

[7] Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., ... & Wooters, C. (2003, April). The ICSI meeting corpus. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on* (Vol. 1, pp. I-364). IEEE.

[8] Klein, A., Maier, E., Maleck, I., Mast, M., & Quantz, J. (1995). *Dialogue acts in VERBMOBIL*. Univ..

[9] Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992, March). SWITCHBOARD: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on* (Vol. 1, pp. 517-520). IEEE.

[10] Kurematsu, A., Akegami, Y., Burger, S., Jekat, S., Lause, B., MacLaren, V., ... & Schultz, T. (2000). VERBMOBIL dialogues: multifaced analysis. In*INTERSPEECH* (pp. 712-715).

[11] Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., ... & Weinert, R. (1991). The HCRC map task corpus. *Language and speech*,*34*(4), 351-366.

[12] op den Akker, H., & Schulz, C. (2008). Exploring features and classifiers for dialogue act segmentation. In *Machine Learning for Multimodal Interaction* (pp. 196-207). Springer Berlin Heidelberg.

[13] Poel, M., Stegeman, L., & op den Akker, R. (2007). A support vector machine approach to dutch part-of-speech tagging. In *Advances in intelligent data analysis VII* (pp. 274-283). Springer Berlin Heidelberg.

[14] Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, *19*(2), 313-330.

[15] Jurafsky, D., Shriberg, E., Fox, B., & Curl, T. (1998, August). Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings of ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers* (pp. 114-120).

[16] Litman, D. J., Hirschberg, J. B., & Swerts, M. (2000, April). Predicting automatic speech recognition performance using prosodic cues. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference* (pp. 218-225). Association for Computational Linguistics.

[17] S. Rosset and L. Lamel, “Automatic detection of dialog acts based on multi-level information,” in icslp, Jeju Island, October 2004, pp. 540–543.

[18] Lendvai, P., van den Bosch, A., & Krahmer, E. (2003). Machine learning for shallow interpretation of user utterances in spoken dialogue systems. In *Proc. of EACL-03 Workshop on Dialogue Systems: interaction, adaptation and styles of management* (pp. 69-78).

[19] Andernach, T. (1996). A machine learning approach to the classification of dialogue utterances. *arXiv preprint cmp-lg/9607022*.

[20] Zimmermann, M., Hakkani-Tür, D., Shriberg, E., & Stolcke, A. (2006). Text based dialog act classification for multiparty meetings. In *Machine Learning for Multimodal Interaction* (pp. 190-199). Springer Berlin Heidelberg.

[21] Petukhova, Volha Viktarauna. "Multidimensional dialogue modelling." *Promotor: H. Bunt. Tilburg* 1 (2011).

[22] Hirschberg, J., & Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. *Computational linguistics*, *19*(3), 501-530.

[23] Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., ... & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, *26*(3), 339-373.

[24] Geertzen, J., Petukhova, V., & Bunt, H. (2007, September). A multidimensional approach to utterance segmentation and dialogue act classification. In*Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, Antwerp*(pp. 140-149).

[25] Reithinger, N., & Klesen, M. (1997, September). Dialogue act classification using language models. In *EuroSpeech*.

[26] Samuel, K., Carberry, S., & Vijay-Shanker, K. (1998, August). Dialogue act tagging with transformation-based learning. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2* (pp. 1150-1156). Association for Computational Linguistics.

[27] Keizer, S. (2003). *Reasoning under uncertainty in natural language dialogue using Bayesian networks*. Twente University Press.

[28] Lendvai, P., Bosch, A. V. D., Krahmer, E., & Canisius, S. (2004). Memory-based robust interpretation of recognised speech. In *9th Conference Speech and Computer*.

[29] Fishel, M. (2006). Dialogue Act Recognition Techniques. *GSLT/NGSLT course on dialogue systems: Linkoping University, Sweden*.

[30] Kim, M. J., Rim, H. C., Park, S. Y., & Han, K. S. (2013). N-gram based Dialogue Act Classification Model with Position Numbering. *Advances in Information Sciences & Service Sciences*, *5*(10).

[31] Manning, C. D. (1999). *Foundations of statistical natural language processing*. H. Schütze (Ed.). MIT press.

[32] Rabiner, L., & Juang, B. H. (1986). An introduction to hidden Markov models.*ASSP Magazine, IEEE*, *3*(1), 4-16.

[33] Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." (2011).

[34] Cohen (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960:37-46.

[35] Fleiss, J. L. (1981). Statistical methods tor rates and proportions. *Nueva York: Wiley*, *8*.

[36] Fišel, M. (2007). Machine learning techniques in dialogue act recognition. *Eesti Rakenduslingvistika Ühingu aastaraamat*, (3), 117-134.

[37] Balkin, S. D., & Ord, J. K. (2000). Automatic neural network modeling for univariate time series. *International Journal of Forecasting*, *16*(4), 509-515.