Dialogue Act Recognition for Text-based Sinhala

Sudheera Palihakkara, Dammina Sahabandu, Ahsan Shamsudeen, Chamika Bandara, and Surangika Ranathunga

Department of Computer Science and Engineering
University of Moratuwa
Katubedda 10400, Sri Lanka
{sudheera.10,dammina.10,ahsan.10,chamika.10,surangika}@cse.mrt.ac.lk

1 Abstract

This paper discusses the application of classical machine learning approaches to the task of Dialogue Act Recognition for text-based Sinhala. A study was carried out to identify a dialogue act tag set for Sinhala. A new corpus using Sinhala subtitles for English movies was created and was annotated with the selected dialogue acts. Evaluation of the dialogue act recognition system was performed using features that were used for English language, plus the newly identified features for Sinhala. Although Sinhala is an under-resourced language without even the basic tools such as a PoS tagger, we managed to achieve good classification accuracy by exploiting Sinhala specific features. As far as we are aware, this is the first research on dialogue act recognition on the family of Indo-Iranian languages.

1 Introduction

Sinhala is the native language of the Sinhalese people, the largest ethnic group in Sri Lanka numbering about 16 million. Considering the other ethnic groups using Sinhala as the second language, Sinhala can be said to be actively used by 19 million people.

Sinhala is the only language that most of the Sinhalese are fluent in. According to the Department of Census and Statistics Sri Lanka, as of 2007, roughly about 50% of the urban youth can read an English newspaper, while in rural areas, this value is well below 40%[1]. In contrast, the overall literacy rate of the country is 98.1%. Therefore there is a dire need for Sinhala language computing. With the implementation of Sinhala Unicode, the platform for this has been set. However the amount of research carried out in the area of Natural Language Processing (NLP) for Sinhala is not adequate. Unlike languages such as English, Spanish or French that are being used by larger populations in the world, Sinhala is restricted to Sri Lanka. This has an adverse impact on the progress made in Sinhala NLP research. Although there exists some preliminary-level research in areas such as Sinhala-English translation (Silva and Weerasinghe, 2008), Sinhala-Tamil (the other official language in Sri Lanka) translation (Sripirakas et al., 2010), and Sinhala spell checking (Jayalatharachchi et al., 2012), the attention paid for processing of spoken and written Sinhala conversations is very low.

The aim of this paper is to lay the first stone to fill this void in processing spoken and written Sinhala conversations. It makes use of the already existing research for Dialogue Act (DA) Recognition for English and explores how it can be used in the context of Sinhala. Given the fact that dialogue act recognition is an important step in understanding spontaneous dialogue, we envisage that this research would pave the path to research in areas of Sinhala NLP such as meeting summarization, question-answering systems, and automated assistance.

As the first step in the process, a corpus was created from Sinhala subtitles for English movies. A set of dialogue acts was identified based

[1] http://www.statistics.gov.lk/MDG/Mid-term.pdf

D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 367–375, Trivandrum, India. December 2015.

on the commonly used dialogue acts for English. Part of the corpus was annotated with these dialog acts. Similarly, feature selection was started with the common features used for English, and later on the study Sinhala-specific features were identified and introduced to improve the classification accuracy. We also experimented with multiple classifiers to select the best performing classifier for Sinhala.

When carrying out Dialogue Act recognition for Sinhala, unavailability of foundational NLP research for Sinhala was a major limitation. For example, Part of Speech (PoS) tags are considered as a successful candidate in the feature set for dialogue act recognition (Verbree et al., 2006). The set of PoS tags has been identified for English and there are many English PoS taggers giving very good accuracy. In contrast, Sinhala PoS tagging is at its inception stage (Herath and Weerasinghe, 2004). Despite these limitations, we managed to achieve a good level of accuracy for Sinhala Dialogue act recognition, by exploiting the Sinhala language-specific features. As far as we are aware, this is the first research on dialogue act recognition on the family of Indo-Iranian languages.

The rest of the paper is organized as follows. Section 2 discusses some important characteristics of Sinhala language and related research in Sinhala language computing. Section 3 gives a brief introduction to the area of dialogue act recognition. Sections 4, 5 and 6 discuss the corpus we created, the dialogue act tag set used in the study, and a discussion on feature selection, respectively. Section 7 presents the results of the study, and finally Section 8 concludes the paper.

2 Sinhala Language and Computing in Sinhala

Sinhala language is more than two thousand years old. It is a language akin to Hindi, Bengali and other north Indian languages. Its closest relative is the language spoken in Maldives islands, Divehi (Pannasara and Arachchi, 2011). Contemporary Sinhala has been influenced by a wide variety of languages including Pali, Sanskrit, Tamil, Portuguese, Dutch and English. Sinhala alphabet is an abugida used in Sinhala writing system, which is a member of Brahmic family script. It is one of the longest alphabets in use today.

As shown in Figure 1, Sinhala belongs to the Indo-Aryan branch of Indo-Iranian language family, which along with Germanic belongs to the larger Indo-European language family. English and German languages are descendants of the Germanic branch.

European family, the Uralic family, the Altaic family, the Sino-Tibetan family, the Afro-Asiatic family and the Niger-Congo family can be considered as the origins of some of the major modern languages (Holman et al., 2011). As depicted in Figure 1, both Sinhala and English languages are descendants of the Indo-European language family.

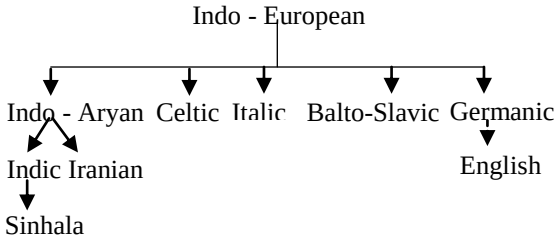

Figure 1. Language Families

Following are some examples on how spoken Sinhala differentiates from spoken English.
In English, the tag question, "isn't it.", "aren't you" or "don't they" agrees with the subject of the sentence that precedes. Its Sinhala equivalent is simply "නේ (ne)?" tagged to the end of the sentence, irrespective of its subject. Some examples are provided in Table 1.

Sentence in Sinhala	**Phonetic Pronunciation**	**English Meaning**
ඔයා වතුර බොනවනේ?	oya wathura bonava, ne?	You drink water, don't you?
අපි වතුර බොමුනේ?	api wathura bomu, ne?	Let's drink water, shall we?
ඔයා ඇමෙරිකන් නේ?	oya aemerikan, ne?	You're American, aren't you?

Table 1. Tag Questions in Sinhala and English

To intensify the meaning of an adjective (such as 'large'), English speakers add 'very' before it: very large. Sinhala speakers have another way of intensifying the meaning of adjectives by lengthening a vowel of the adjective itself. Thus the term 'ලොකු (loku)' (large) can be made into 'ලොකූ (lokuuu)' (very large).

Although there exists much dissimilarity between Sinhala and English, it is not difficult to identify some similarities between the two languages through a much closer inspection.

If we consider the phonetic pronunciation of different words, we can observe similarities in languages of the Indo-European family. For an example the English word month pronounced in German as *Monat*, in Welsh as *mis*, in Italian as *mese* and in Sinhala as *masaya*[2].

Moreover, the set of punctuation marks used in both Sinhala and English are identical, although the ancient Buddhist scriptures did not use such punctuations. This could be due to the influence that Colonial English had over Sinhala.

As mentioned earlier, there exists some preliminary NLP research for Sinhala such as a Sinhala PoS tagger (Herath and Weerasinghe, 2004), a Sinhala WordNet (Wijesiri et al., 2014), English-Sinhala translation (Silva and Weerasinghe, 2008), and Sinhala-Tamil translation (Sripirakas et al., 2010). However, none of this work can be said to be comprehensive or completed. Still there is a considerable amount of work to be done in implementing the basic NLP tools for Sinhala.

3 Dialogue Act Recognition

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 modelling a spontaneous dialogue have already been identified. Among these clearly identified characteristics, "Dialogue Acts" hold an important place.

3.1 Speech Acts and Illocutionary Forces

A speech act in linguistics is an utterance that has performative function in language and communication (Searle, 1985). 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 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.

Dialogue acts are a specialized version 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 such as yes-no-questions, and open questions. 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 depend on the application.

Austin (1975) 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?, isn't 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.

3.2 Process of Identifying Dialogue Acts

The process of identifying the Dialogue Act of an utterance in a particular language consists of a fixed set of steps (Král and Cerisara, 2012). This process is independent of 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. This is a very critical step since the accuracy of identifying the Dialogue Acts heavily depends on the identified feature set. Then DA models will be trained on these identified features set. The trained DA model can be used to determine the dialogue act of a given utterance. To make the process of dialogue act recognition easier, segmentation of the dialogues

[2]Used Google translator.

into utterances needs to be carried out independently, or alternatively realized during the recognition step with joint DA recognition and segmentation models.

4 Corpus

Since no corpus was available for dialogue act recognition for Sinhala, it was required to build a standard corpus from the scratch. We identified several approaches for this task:

1. Translate an existing standard English corpus
2. Collect written (typed) conversations from a Sinhala chat tool
3. Collect Sinhala subtitles in English movies
4. Collect conversations from Sinhala novels.
5. Collect telephone conversations carried out in Sinhala.

Among above approaches, the last one is a very difficult task to perform, because there is no solid research for speech-to-text conversion of Sinhala. Collecting conversations from Sinhala novels was found to be not possible due to the public unavailability of novels in digital format. Finding translators was not possible so we abandoned the first option.

Then we deployed a Sinhala chat tool for public use and collected conversations. At the beginning, this approach seemed promising but the process was slow because it was difficult to get volunteers. Moreover, volunteers tend to use English words in the middle of Sinhala utterances. Also they used slangs and urban words more often, which makes the classification more complex. Although we understand that a dialogue act recognition system should accept the existence of such non-standard words, this was considered out of scope for the current research.

Then we extracted utterances from Sinhala subtitles of English movies. The translation of English movies is a result of a community-based crowd sourcing effort. About 10 full-time translators are contributing to this under the trade name of "baiscope.lk"[3]. In Sri Lanka, there is a large population that enjoys Hollywood movies and TV series. However, their low English literacy is a problem when understanding these movies and TV series. The aim of baiscope.lk is to provide Sinhala subtitles for English movies and TV series. The subtitle creation process is governed by a set of rules and regulations. The subtitles are almost in grammatically correct Sinhala.

One issue with this method is that some movies have frequent scene changes. This is problematic for extracting consistent conversations. To overcome this we had to manually select the movies that contained long consistent scenes. We collected about 1.8 million utterances using this method for 2306 movies. A common characteristic of these selected movies was that they involved realistic characters and real life situations. Therefore the conversations taking place in these movies are general human conversations that do not refer to any specific domain (e.g. war). This is exactly what we required, in order to carry out dialog act recognition in general human conversations.

Extraction and segmentation of utterances were done in a semi-automatic manner to build a more conversation-oriented corpus. Extracting the utterances from a subtitle file consists of several steps. First step is to omit the time-related information mentioned alongside utterances. Then the filtering out of advertisements and symbolic characters takes place. Finally improperly used punctuation marks are removed. These include the use of multiple exclamation/question marks in order to emphasize the emotion conveyed in the movie scene instead of using just one right after an utterance. Segmentation is done manually by checking each line, because one statement is sometimes broken into few lines in the subtitles due to a scene change in the middle of an utterance in the movie. If any such lines are found, they can be combined into a single line.

The final "Sanwada" corpus contains 1.8 million utterances including tagged 12,000 utterances. In Sinhala, the term "Sanwada" means conversation.

5 Tag Set

There exists many research related to dialog act tag sets, and dialog act annotation (Bunt et al., 2010, Bunt et al., 2012). To select a suitable tag set for Sanwada corpus, we adapted a generic tag set by referring to the DAMSL (Allen and Mark, 2013) tag set and the study by Stolcke et al., (2000). To measure the necessity and sufficiency of this tag set for tagging Sanwada corpus, we performed several iterations of manual tagging

[3]http://www.baiscopelk.com/

for a separate a set of samples. These samples were chosen from a set of tagged utterances that were not included in the "Sanwada" corpus. In each iteration, we added necessary new tags and removed unnecessary tags from the set. Table 2 lists the final tag set along with the percentage of occurrence in the manually tagged Sanwada corpus.

Dialogue Act Tag	Percentage
Statement	48.51%
Yes-No Question	12.87%
Request/Command/Order	10.23%
Open Question	9.78%
Back-channel/Acknowledge	7.39%
Conventional Opening	2.58%
Backchannel Question	2.31%
No Answer	1.42%
Yes Answers	1.36%
Apology	1.33%
Thanking	0.75%
Opinion	0.44%
Aadoned/Uninterpretable/Other	0.44%
Conventional Closing	0.31%
Expressive	0.17%
Reject	0.11%

Table 2. Selected Dialogue Act Tag Set

Wh-Question is one of the major tags used in related work (Stolcke et al., 2000). The presence of 'WH' letters as in 'what', 'when', 'why', 'which' etc. in an utterance is used as a feature in order to identify Wh-Questions. But considering the lexical characteristics of Sinhala this tag is not applicable. For example, in Sinhala, 'මොකක්ද' means 'what', 'කීයටද' means 'when', and 'ඇයි' mans 'why'. As can be seen, the first character of these Sinhala words is different in each word, as opposed to the English words. Therefore we used more generic tag Open-Question for questions in general unless it is a Yes-No Question or a Backchannel Question.

In the initial tag set we had two separate tags for Request and Command/Order. For English there is a clear separation in utterances between these two tags. Most of the Requests include the word "please" or a similar phrase in contrast to Command/Orders where it does not. In Sinhala, different forms of the same word are used to indicate whether it is a request or a command. For example, වහන්න (wahanna) is used in requests in a polite manner to say close something (e.g. a door) where වහපං (wahapan) is used in orders.

It should also be noted here that English-Sinhala translation in baiscope.lk is not just a mere one-to-one mapping from English to Sinhala. This is because the translation process is subjective. The translators generate subtitles while watching the movie. Therefore they capture the prosodic and other contextual information in the Sinhala subtitles to a great extent. For example, consider a movie scene where an actor asks another actor to "close that door" in a very harsh tone. The corresponding Sinhala subtitle uses command-type words "දොර වහපං" (dora wahapan) instead of request-type words "දොර වහන්න" (dora wahanna).

The rate of occurrence of Backchannel Questions is comparatively high in Sinhala. Therefore we introduced it as a separate tag. Backchannel questions are Back-Channels or Acknowledges in question form. For example, in Sinhala conversations we often come across the phrase "එහෙමද?" (ehemada?) in response, roughly it means "is it?". It should also be noted that there is no relevant Sinhala phrase for the commonly used English term "isn't it".

To tag the Sanwada corpus using the tags listed in Table 2, we have selected four independent contributors. After tagging the complete corpus manually, we have calculated the inter-annotator agreement among them using Fleiss kappa (Fleiss, 1981) value and the agreement was 0.8161. To calculate the kappa value we implemented a tool based on the equations introduced by Fleiss (1981).

6 Feature Selection

Our target was to test the performance of features already identified in related work for English and distinguish the relevant features for Sinhala. Also we have identified several new features exclusive for Sinhala.

6.1 Identified features from related work

We have identified 14 features that can be used in textual dialogue act recognition from previous studies (Verbree et al., 2006; Rosset and Lamel, 2004). Among those 14 features we selected only 7 features for our study considering the applicability to Sinhala and other few concerns that are discussed below. Table 3 lists these 14 features along with their selection status.

Since we are using Bigrams as a feature, feature 8 and 9 were omitted. Feature 10 is omitted due to the unavailability of a Sinhala PoS tagger. Taking previous Dialogue Acts as a feature can introduce a cumulative error as described by Lendvai et al., (2003). Unigrams are ineffective

for long utterances, although their effectiveness has been shown for chat messages (Ivanovic, 2008).

Feature	Status
1. Number of words in the segment	Selected
2. Bigrams/Trigrams of words	Selected
3. Previous Dialogue Act	Selected
4. Verb of the Sentence	Selected
5. Punctuation marks	Selected
6. Grammar pattern	Selected
7. Frequent words for each tag	Selected
8. First two words	Not-selected
9. Last two words	Not-selected
10. First verb type/ Second verb type	Not-selected
11. Words in last 10 Dialogue Acts	Not-selected
12. N-grams of previous Dialogue Acts	Not-selected
13. Bag-of –words	Not-selected
14. Unigrams	Not-selected

Table 3. Selected Features

6.2 Exclusive features for Sinhala

Last letter of the last word of the utterance is one feature that we have identified. Unlike in English, the last letter of the utterance makes a big impact on the dialogue act of the utterance. For instance, most of the Yes/No questions end with the letter 'ද'(da), most of Request/Command/Order ends with one of the letters 'න්'(n), 'න'(na), or 'නු'(nu), and most of the open questions end with 'නේ'(ne). Not only the last letter but also the last word of an utterance is an exclusive feature for Sinhala.

The presence of specific Sinhala cue phrases is another identified feature. Table 4 lists some identified cue phrase sets.

6.3 Identified Features

Next follows all the major features that could be used for dialogue act recognition.

1. *Cue Phrases*: presence of connective expressions.
2. *Number of words in the segment*: self-explanatory.
3. *Bigrams/Trigrams of words*: Adjacent two words in an utterance is considered as a bigram, likewise trigram is adjacent three words.
4. *Previous Dialogue Act*: The dialogue act of the previous utterance.
5. *Verb of the Sentence*: self-explanatory
6. *Punctuation marks*: The appearance of the question mark, exclamation mark, Full stop, etc. in the utterance. In Sinhala same punctuation marks are used as in English.
7. *Grammar pattern*: The Sinhala grammar pattern(s) of the sentences in the utterance.
8. *Last word of the utterance*: self-explanatory.
9. *Frequent words for each tag*: For each tag the most frequent words appear in the training set of utterance.
10. *End letter of the last word of the sentence*: self-explanatory.

Sinhala cue phrase(s)	Phonetic Pronunciation	English cue phrase
ඇත්තෙන්ම	aeththenma	actually
සහ, හා	saha, haa	and
නිසා, හින්ද	nisa, hinda	because
එසේම	esema	also
එහෙත්, නමුත්	eheth, namuth	but
වගේ, වැනි, වාගේ	wage, waeni, waage	like
ඉතින්, එවිට	ithin, ewita	then
හෝ	ho	or
හරි	hari	well
එනිසා, එබැවින්	enisaa, ebawin	so

Table 4. Cue Phrases

6.4 Feature Selection Experiments

The idea of the experiments is to identify the most contributing features for classifying and the most effective combinations of the features. From the aforementioned 10 features, 8 were selected based on the performance evaluation (with 10 features, it is computationally expensive than for 8 features to go through all possible combinations)[4].

[4] For 10 features have to go through 2^{10} i.e. 1024 combinations where for 8 features it's only 2^8 i.e. 256.

We used WEKA (Hall et al., 2009) toolkit for classification. To achieve above described task we used the InfoGain Attribute Evaluator of WEKA and obtained the InfoGain values. Table 5 displays the results. The InfoGain value evaluates the worthiness 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. However this does not mean that the use of that feature alone is able to conduct the entire classification.

Rank	Feature	InfoGain
1	Punctuation marks	0.71
2	Last word of the utterance	0.60
3	Frequent words for each tag	0.42
4	Trigrams/Bigrams	0.31
5	Last letter of the last word of the sentence	0.30
6	Verb of the Sentence	0.24
7	Number of words in the segment	0.18
8	Cue Phrases	0.17

Table 5. Individual Feature Performance

From this result set we can observe that the most contributing feature for the task is punctuation marks. The processed subtitles that we used have been properly written with the use of punctuation marks. This particular feature has been effective in distinguishing questions (Open Question, Yes/No Questions and Back-channel Questions) from other tags. Some of the features that we identified as exclusive features for Sinhala (last word of the utterance and last letter of the utterance) also contribute a considerable amount.

Feature 3 in the table (Frequent words for each tag) keeps track of the most frequent words used in the entire corpus and uses the presence of those words in a particular utterance as a feature for the classifier. For this task we used WEKA's StringToWordVector option with the word count of 100. This feature has not been widely used in related work but we could observe that this feature works well.

There were limitations on finding the verb of the sentence precisely due to the lack of resources for PoS tagging for Sinhala. Therefore we used a set of commonly used Sinhala verbs and checked the presence of those verbs in a given utterance as feature.

From the above mentioned features we have selected the best performing six features listed in the Table 6 by testing all the combinations of features on the J48 WEKA classifier.

Feature
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

Table 6. Best Performing Features

For the 8 different features there are 256 different combinations of feature sets. We went through all these different combinations and classified them using a trained J48 classifier. The features mentioned in Table 6 yielded the maximum accuracy on the testing set. This feature set achieved F-measure value of 0.755 with a precision 0.788 and recall 0.755.

7 Results

For classification task we have used 8000 utterances as training set and 4000 utterances as testing set. Each entry was labeled with exactly one dialog act. As the first step we have tested the classification accuracy by just using the features used for dialogue act recognition in English. From the best performing features stated in Table 6, Punctuation marks, Trigrams/Bigrams and Frequent words for each tag are the three features used in the related work. The other three features are specific for Sinhala. Using those three features used for English we were able to gain an accuracy of 71.14% in classification using the J48 classifier. Then we have used all six features and classified using the same classifier and we were able to improve the accuracy to 78.68%.

As the next step we have used the same feature set and classified the same data set using different classifiers to model the performance of different classifiers on Sinhala. Table 7 lists the classifiers in the descending order of F-measure value. F-measure represents a value of accuracy of the tests performed, which is calculated using recall and precision values. We can observe that RandomForest, SimpleLogistic and LMT classifiers give the highest F-measure.

Classifier	Recall	Precision	F-measure
RandomForest	0.792	0.780	0.776
SimpleLogistic	0.794	0.772	0.765
LMT	0.794	0.760	0.762
PART	0.786	0.757	0.756
J48	0.789	0.773	0.755
NaiveBayes	0.756	0.728	0.732
REPTree	0.782	0.761	0.731
DecisionTable	0.761	0.757	0.727
SMO	0.769	0.728	0.708
DecisionStump	0.639	0.413	0.501
HoeffdingTree	0.677	0.522	0.577

Table 7. Classifier Performance

8 Conclusion

This paper explored how Dialog Act recognition can be carried out for Sinhala, which is an under-resourced language. We built a corpus using Sinhala movie subtitles, and defined a suitable dialogue act tag set for this corpus based on the results of a few tests performed on the corpus. The experiments done on the corpus for recognizing dialogue acts obtained reasonable results and showed that Sinhala-specific features can be used to improve Sinhala dialogue act recognition.

The feature selection test explored new ways of extracting information from the utterances and we identified a best performing feature set for the Sinhala Language. Despite the small size of the feature set, we managed to achieve a reasonable accuracy in different classifiers. The classifier tests revealed that most of the classifiers perform well with the Sinhala corpus without any classifier parameter tuning. We reached to 78.68% accuracy of dialogue act tagging with RandomForest classifier.

As future work, we suggest taking lower level information such as prosody into the picture and defining features related to it. Since we have a very large unlabelled data set, it is possible to explore the use of unsupervised learning techniques for dialog act recognition for Sinhala. We also envisage that this research would pave the path for more Sinhala related research such as meeting summarization in Sinhala and Sinhala question answering systems.

Reference

JM Allen and Mark Core. 1997. draft of DAMSL: Dialog act markup in several layers.

John Langshaw Austin. 1975. How to do things with words, volume 367. Oxford university press.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee et al. 2010 Towards an ISO standard for dialogue act annotation. In Seventh conference on International Language Resources and Evaluation.

Harry Bunt, Michael Kipp and Volha Petukhova. 2012. Using DiAML and ANVIL for multimodal dialogue annotations. In proceedings of the Language Resources and Evaluation Conference.

Joseph L Fleiss. 1981. The measurement of interrater agreement. Statistical methods for rates and proportions, 2:212–236.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1):10–18.

Dulip Lakmal Herath and Ruvan Weerasinghe. 2004. A stochastic part of speech tagger for Sinhala. In Proceedings of the 06th International Information Technology Conference, pages 27–28.

Eric W Holman, Cecil H Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, et al. 2011. Automated dating of the worlds language families based on lexical similarity. Current Anthropology, 52(6):841–875.

Edward Ivanovic. 2008. Automatic instant messaging dialogue using statistical models and dialogue acts. Master's thesis, University of Melbourne.

Eranga Jayalatharachchi, Asanka Wasala, and Ruvan Weerasinghe. 2012. Data-driven spell checking: The synergy of two algorithms for spelling error detection and correction. In International Conference on Advances in ICT for Emerging Regions, pages 7– 13. IEEE.

Pavel Král and Christophe Cerisara. 2012. Dialogue act recognition approaches. Computing and Informatics, 29(2):227–250.

Piroska Lendvai, Antal van den Bosch, and Emiel Krahmer. 2003. Machine learning for shallow interpretation of user utterances in spoken dialogue systems. In Proceedings of the EACL-03 Workshop on Dialogue Systems: Interaction, Adaptation and Styles of Management, pages 69–78.

Okampitiye Pannasara and V. Arachchi. 2011. Sinhala bhasha vikashaya saha shilalekhana wimarshana. Colombo.

Sophie Rosset and Lori Lamel. 2004. Automatic detection of dialog acts based on multi-level information. pages 540–543.

John R Searle. 1985. Expression and meaning: Studies in the theory of speech acts. Cambridge University Press.

Anne Mindika Silva and Ruvan Weerasinghe. 2008. Example based machine translation for English-Sinhala translations. In Proceedings of the 09th International IT Conference, pages 27–28.

Sakthithasan Sripirakas, Ruvan Weerasinghe, and Dulip L Herath. 2010. Statistical machine translation of systems for Sinhala-Tamil. In International Conference on Advances in ICT for Emerging Regions, pages 62–68. IEEE.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational linguistics, 26(3):339–373.

Daan Verbree, Rutger Rienks, and Dirk Heylen. 2006. Dialogue-act tagging using smart feature selection; results on multiple corpora. In Spoken Language Technology Workshop, pages 70–73. IEEE.

Indeewari Wijesiri, Malaka Gallage, Buddhika Gunathilaka, Madhuranga Lakjeewa, Daya C Wimalasuriya, Gihan Dias, Rohini Paranavithana, and Nisansa de Silva. 2014. Building a WordNet for Sinhala. In Seventh Global Wordnet Conference, pages 100–108.