Dialogue Act Recognition for Text Based Sinhala

Abstract

This paper is concerned with the classical machine learning approaches to the task of Dialogue Act Recognition for text based Sinhala. A new annotated corpus for Sinhala language is built along with a comprehensive dialogue act tag set. We performed an evaluation based on features identified in utterances. Evaluation is performed on features that used in other studies as well as newly identified features for Sinhala. Using the best performing feature set we have performed an evaluation for effectiveness of classifiers. Considering the test results on features, we identified the best performing feature set for Sinhala language. Considering the result of classifier test, we identified the best performing classifier for Sinhala language.

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, English literacy rate of the population in 2001 is 13.63%, while the Sinhala literacy rate[1] is 91%. 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, Sinhala-Tamil (the other official language in Sri Lanka) translation, Sinhala spell checking [12], the attention paid for processing of spoken Sinhala is very low.

The aim of this paper is to make use of the already existing research for Dialog Act Recognition for English and explore how it can be used in the context of Sinhala. Given the fact that dialog act recognition is an important step in understanding spontaneous dialogue, we envisage that this research would pave the path to research in areas such as meeting summarization, question-answering systems, and automated assistance.

A corpus was created from Sinhala subtitles for English movies. A set of dialog acts was identified based on the commonly used Dialog acts for English. Similarly, feature selection was started with the common features used for English, and later on the study Sinhala-specific features were identified to improve the classification accuracy. We also experimented with multiple classifiers to select the best performing classifier for Sinhala.

However, none of these are considered to be completed. When carrying out Dialog Act recognition for Sinhala, unavailability of foundational NLP research for Sinhala was a major limitation. For example, PoS tags are considered as a successful candidate in the feature set for dialog act recognition [12]. 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 [4]. Despite these limitations, we managed to achieve a good level of accuracy for Sinhala Dialog act recognition, by exploiting the Sinhala language-specific features. As far as we are aware, this is the first research on dialog 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, 4 and 5 discuss the corpus we created, the dialog act tag set used in the study, and a discussion on feature selection, respectively. Section 6 presents the results of the study, and finally Section 7 concludes the paper.

[1] Population aged 10 years and over

2. Sinhala Language and Computing in Sinhala

Sinhala language is more than two thousand years old and it is a language akin to Hindi, Bengali and other north Indian languages. Its closest relative is the language spoken in Maldives islands, Divehi [8]. 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.

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.

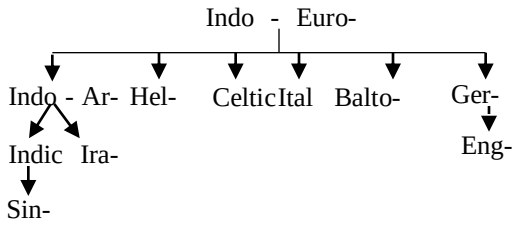

Figure 1 Language Families.

When considering language families, the Indo-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 [5]. As depicted in Figure 1, both Sinhala and English languages are descendants of Indo-European language family.

Following are some examples on how Sinhala differentiates from 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 තේ බොනවනේ?	oya te bonava, ne?	You drink tea, don't you?
අපි තේ බොමුනේ?	api te bomu, ne?	Let's drink tea, shall we?
ඔයා ඇමෙරිකන් නේ?	oya aemerikan, ne?	You're American, aren't you?

Table 1 Tag Questions.

Another example is 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 'loku' (large) can be made into 'lokuuu' (very large).

Although there exist many dissimilarities 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. This could be due to the influence that Colonial English had over Sinhala.

As mentioned earlier, there exist some preliminary NLP research for Sinhala. [13] [4]

3. Corpus

Since no corpus was available for dialog act recognition for Sinhala, it was required to build a standard corpus from the scratch. We tried out three approaches for this task.

1. Translate an existing standard English corpus
2. Sinhala chat tool
3. Sinhala movie subtitles

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 and the volunteers were tend use to

[2] Used google translator

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 dialog act recognition system should accept the existence of such non-standard words, this was considered out of scope for the current research.

Then we tried to extract utterances from Sinhala subtitles of English movies. The translation of English movies is a result of a community-based crowed sourcing effort. About 10 full-time translators are contributing to this under the trade name of "baiscopelk"[3]. In Sri Lanka, there is a large population that enjoy 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. 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.

Extraction and segmentation of utterances were done manually to build a more conversation-oriented corpus. Extracting the utterances from a subtitle file consist 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 eliminating the improperly used punctuation marks. Such as using multiple exclamation/question marks instead of using one right after an utterance in order to emphasize the emotion conveyed in the movie scene. Segmentation is done manually by checking each line for one statement broken in to few lines in the subtitles. It is a result of a scene change in the middle of an utterance in the movie. If any such lines found, can combine them into a single line.

The final corpus contains 1.8 million utterances including tagged 12,000 utterances. It is publicly available[4] under the name of "Sanwada". In Sinhala, the term "Sanwada" means conversation.

4. Tag Set

To select a suitable tag set for Sanwada corpus, we adapted a generic tag set by referring to the DAMSL [1] tag set and the study by Stolcke et al. To measure the necessity and sufficiency for tagging Sanwada corpus we performed several iterations of manual tagging for a separate set samples. These samples are 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.

Dialog 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%
Abadoned/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 [10]. 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. So 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

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

[4] http://web.sanwada.robotics.lk/download.html

"Please" or a similar phrase in contrast to Command/Orders where it does not. In Sinhala, different forms of the same word is 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 (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 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 වහන්න" (dora wahapan) instead of request-type words "දොර වහපං" (dora wahanna).

The rate of occurrence of Backchannel Questions are comparatively high in Sinhala. So 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?".

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 [2] value and the agreement was 0.8161. To calculate the kappa value we implemented a tool based on the equations introduced by Fleiss [2].

5. Feature Selection

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

1.1 Identified features from related work

We have identified 14 features that can be used in textual dialogue act recognition from previous studies [12] [9]. 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 features along with their selection status.

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.

In

Table 3, since we are using the Bigrams as a feature, feature 8 and 9 were omitted. Feature 10 is omitted due to the unavailability of Sinhala PoS tagger. Taking previous Dialogue Acts as features can introduce a cumulative error as described by Lendvai [7]. Unigrams are ineffective for long utterance, although their effectiveness has been shown for chat messages [6].

1.2 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 ends with the letter 'ද'(da), most of Request/Command/Order ends with one of the letters 'න'(n), 'න'(na), or ' නු'(nu), most of 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.

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.

1.3 Identified features

Next follows all the major features used for dialog 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.

1.4 Feature set selection

The idea of the experiment 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. Because with 10 features, it is computationally expensive than for 8 features to go through all possible combinations.[5]

We used WEKA [3] Java library for classification. To achieve above described task we used InfoGain Attribute Evaluator of WEKA and obtained the InfoGain values.

Table 5 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.

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. In the subtitles that we used has 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 contributes a considerable amount.

Frequent words for each tag feature keeps track of the most frequent words used in the entire corpus and uses the

[5] 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

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 can observe that this feature works well.

There were limitation on finding the Verb of the sentence precisely such as lack of resources for PoS tagging for Sinhala. Therefore we used a set of commonly used Sinhala verbs to check 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 the all combinations of features on a selected 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

We have used the J48 WEKA classifier to perform this test. 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 feature mentioned in the

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.

6. Results

For classification WEKA uses various classifiers according to the methods discussed in Section II. We used selected best performing feature set with available classifiers and compared the accuracy values and their performance.

For classification task we have used 8000 utterances as training set and 4000 utterances as testing set. 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 the

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.

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

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 value. We can observe that SimpleLogistic and LMT classifiers gives the highest recall value. That means they have identified more correctly tagged utterances compared to other classifiers.

7. Conclusion

In this paper we discussed about adapting the classification techniques to Sinhala Language. We built a corpus using Sinhala movie subtitles, and defined suitable dialogue act tag sets for Sanwada corpus based on the results of a few tests performed on the corpus. The experiments done on Sanwada corpus for recognizing dialogue acts obtained reasonable test results and showed that specific features for Sinhala can be used for improve Sinhala dialogue act recognition. Our work was carried out using a relatively small number of tag sets and features, so despite the fact we achieved reasonably good results, the performance can be increased further more. 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. 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 in WEKA.

As future work, we suggest taking lower level information like prosody in to the picture and defining features on it. Classifier optimization is another aspect we have not covered in our study.

References

[1] Allen, J., and G. Mark. "Core. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers." (2013).

[2] Fleiss, Joseph L., Bruce Levin, and Myunghee Cho Paik. "The measurement of interrater agreement." Statistical methods for rates and proportions 2 (1981): 212-236.

[3] Hall, Mark, et al. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11.1 (2009): 10-18.

[4] Herath, Dulip Lakmal, and A. R. Weerasinghe. "A Stochastic Part of Speech Tagger for Sinhala." Proceedings of the 06th International Information Technology Conference (IITC'04), Colombo, Sri Lanka. 2004.

[5] Holman, Eric W., et al. "Automated dating of the world's language families based on lexical similarity." Current Anthropology 52.6 (2011): 841-875.

[6] Ivanovic, Edward. "Automatic instant messaging dialogue using statistical models and dialogue acts." University of Melbourne, Department of Computer Science and Software Engineering, 2008.

[7] Lendvai, Piroska, Antal van den Bosch, and Emiel Krahmer. "Machine learning for shallow interpretation of user utterances in spoken dialogue systems." InProc. of EACL-03 Workshop on Dialogue Systems: interaction, adaptation and styles of management, pp. 69-78. 2003.

[8] Paññāsāra, O. and Āracci, V. (2011). Siṃhala bhāṣā vikāsaya saha śilālēkana vimarśana. [Koḷamba: Okkampīṭiyē Paññāsāra Himi].

[9] Rosset, Sophie, and Lori Lamel. "Automatic Detection of Dialog Acts Based on Multi-level Information Ý." (2004).

[10] Stolcke, Andreas, et al. "Dialogue act modeling for automatic tagging and recognition of conversational speech." Computational linguistics 26.3 (2000): 339-373.

[11] Verbree, Daan, Rutger Rienks, and Dirk Heylen. "Dialogue-act tagging using smart feature selection; results on multiple corpora." Spoken Language Technology Workshop, 2006. IEEE. IEEE, 2006.

[12]Wasala, Ruwan Asanka, et al. "An Open-Source Data Driven Spell Checker for Sinhala." International Journal on Advances in ICT for Emerging Regions (ICTer) 3.1 (2011): 11-24.

[13] Wijesiri, Indeewari, et al. "Building a WordNet for Sinhala."