MACHINE LEARNING TECHNIQUES WITH NLP FOR DIALOG ACT CLASSIFICATION

J.R.K.C Jayakody*
Department of Computing and Information System, Faculty of Applied Sciences, Wayamba University of Sri Lanka
**Corresponding author (email: kithisrij@wyb.ac.lk)*

Introduction

Dialog act classification based on utterances plays an indispensable role in the dialogue phenomena [1]. The idea of interpreting dialogue behavior in terms of communicative actions such as statements, questions, promises, requests, and greetings goes back to speech act theory which has been an important source of inspiration for modern dialogue systems. Performance of speech systems and automatic machine translation are mainly based on the identification and classification of different dialog acts. Therefore this research is mainly focused on using statistical and basic word features and extracting the features with Natural Language Processing (NLP) techniques to improve the classification accuracy with machine learning techniques. In this study, Decision trees, Naive Bayes classifier, Bayes Network classifier, IBK classifier, Logiboost classifier and Naïve Bayes tree classifier were tested with different features to improve the accuracy.

Methodology

The present research was mainly based on extracting the most appropriate features which are not prosodic. The dataset which was used for the experiment consists of 5778 utterances of text-based human conversations. This data has been recorded from a role-playing game in a 'virtual hospital'. The data correspond to human conversations recorded from the chat channels of this virtual world. There are several doctors, patients and sometimes relative of the patient. The doctor's task is to diagnose the illness of the patient. The conversation data have been manually tagged with the corresponding dialog acts. Altogether, there are 14 dialog acts – Accept, Apology, bye, Continuer, Emotions, Greet, No-Answer, Other, Request, Statement, Thank, Wh-Question, Yes-Answer and Yes-no-Question.

NLP procedures were written to extract basic statistical features as well as word features. Further regular expression grammar rules were used to extract utterance patterns. Finally Weka tool was used for the classification. Experiment was conducted to identify the best feature set. Number of classification algorithms was tested with the features to identify the best classification algorithm.

Results and Discussion

In this research special words were identified specific to each classification type [Table 1]. There is a high probability that if those words are appearing, the classification type would be the identified type. Further, statistical features were used to identify the classification type such as start word, end word, first two words, No of symbols and the length of the question etc.

Table 1: Specific words related to each classification type

Classification Type	Specific Word
Accept	YES, YEP, YEA, YUP
nAnswer	NOPE, NO, NEVER, NOT, NAAH
Greet	?, COULD, HELLO
Bye	SEE YOU,SEE YA, BYE, LEAVE, LOGGING OUT, C U, TATAA, CIAO
Apology	SORRY
Accept	OH, GREAT, OK, EXCELLENT, FINE, SURE, LIKE, GOOD,TRUE, OFCOURSE,COOL
Emotions	LOL, AHHA, HAHA, OOPS, WELL, HEE
Greeting	HI, GOOD MORNING, HEY, NICE TO MEET
Thank	THANKS, THANK YOU, THKS
Request	CAN + SUBJECT, TELL +OBJECT , WOULD+ SUBJECT,MAY+ SUBJECT, SHOULD+ SUBJECT, SALL+ SUBJECT, PLEASE, + SUBJECT'D
Statement	SO, AFTER, AND, ALSO, TO, ALTHOUGH, THEN, UNLESS, USUALLY, BECAUSE, COS, DESPITE, EXCEPT, FOR, FROM , IN, LAST, JUST, OR, OTHERWISE,WHEREAS, WITH,WHILE etc.

NLP Grammar patterns

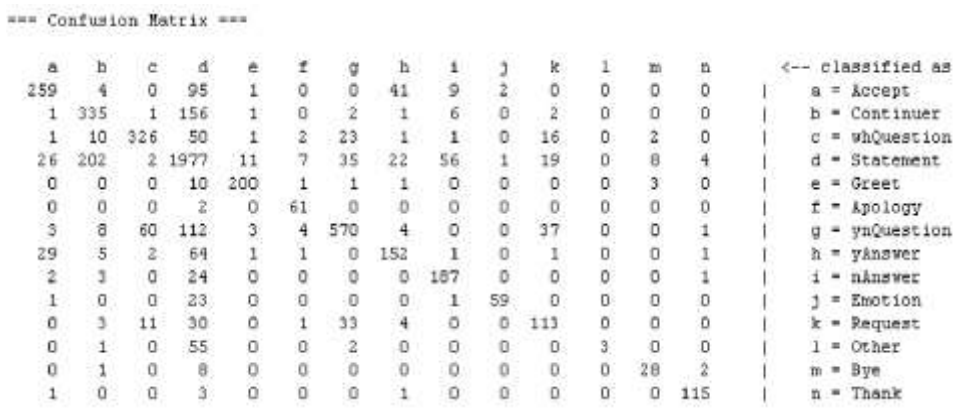
This stage consists of two steps. First, utterances were used to preprocess with NLP techniques. Paragraphs were broken into sentences. Next, taggers were used to tag each word. Rules were generated that made up chunk grammar to identify tag patterns. A tag pattern is considered as a sequence of part-of-speech tags which delimited using angle brackets. e.g. <DT>?<JJ>*<NN>. To find the chunk structure for a given sentence, grammar was developed [Table 2] appropriately and later it was used to create the regular expression parser to identify utterances types.

Table 2: Grammer pattern rules

Rule No	Grammar rule
1	*NP: {<DT>?<JJ>*<NN>}*
2	*NP: {<DT|PP\$>?<JJ>*<NN>}, {<NNP>+},*
3	*{<V.*> <TO> <V.*>}'*
4	*: NP: {<DT|JJ|NN.*>+}* *PP: {<IN><NP>}* *VP: {<VB.*><NP|PP|CLAUSE>+$}* *CLAUSE: {<NP><VP>}*

WEKA was used with different features to improve the classification accuracy [2]. Decision trees, Naive Bayes classifier, Bayes Network classifier, IBK classifier,

Logiboost classifier, NB tree classifier were tested with different feature set to improve the accuracy. Accuracy of the machine learning algorithms was improved based on different features that were generated. Figure 1 is shown the confusion



```
=== Confusion Matrix ===

   a   b   c    d   e   f   g   h   i   j   k   l   m   n      <-- classified as
 259   4   0   95   1   0   0  41   9   2   0   0   0   0   |   a = Accept
   1 335   1  156   1   0   2   1   6   0   2   0   0   0   |   b = Continuer
   1  10 326   50   1   2  23   1   1   0  16   0   2   0   |   c = whQuestion
  26 202   2 1977  11   7  35  22  56   1  19   0   8   4   |   d = Statement
   0   0   0   10 200   1   1   1   0   0   0   0   3   0   |   e = Greet
   0   0   0    2   0  61   0   0   0   0   0   0   0   0   |   f = Apology
   3   8  60  112   3   4 570   4   0   0  37   0   0   1   |   g = ynQuestion
  29   5   2   64   1   1   0 152   1   0   1   0   0   1   |   h = yAnswer
   2   3   0   24   0   0   0   0 187   0   0   0   0   1   |   i = nAnswer
   1   0   0   23   0   0   0   0   1  59   0   0   0   0   |   j = Emotion
   0   3  11   30   0   1  33   4   0   0 113   0   0   0   |   k = Request
   0   1   0   55   0   0   2   0   0   0   0   3   0   0   |   l = Other
   0   1   0    8   0   0   0   0   0   0   0   0  28   2   |   m = Bye
   1   0   0    3   0   0   0   1   0   0   0   0   0 115   |   n = Thank
```

matrix that was generated with WEKA tool for the Logiboost classification algorithm.

Figure 1 : Confusion matrix of Logiboost Classification algorithm

Conclusion and recommendations

According to the research Logiboost were identified as the best classifier for Dialog act recognition. Logiboost algorithm accuracy was improved from 64% to 76% with the improvement of features, especially with the chunk grammar rules. Moreover, statistical features such as first verb type, second verb type were helped immensely to improve the accuracy.

Some of the classifiers were not tested due to the insufficient memory space. SMO classifier had been taken more than 10 hours to give the find outcomes. Therefore it was identified as the least perform classifiers. Out of classification families, Naïve Bayes family gave higher performance compare to other classifiers. All components of the model were automatically trained. Classification accuracies achieved so far are highly encouraging relative to the inherent difficulty of the task as measured by human.

References

[1] M. Mast *et al*., Dialog act classification with the help of prosody. *In Proceedings of the Fourth International Conference on Spoken Language, ICSLP 96, 1996,* Vol. 3, pp. 1732-1735.

[2] S. Singhal, M. Jena. (2013). A study on WEKA tool for data preprocessing, classification and clustering. *International Journal of Innovative Technology and Exploring Engineering,* 2(6), pp. 250-253, 2013.