Dialogue Act Recognition using Cue Phrases

Jun Araki
Computer Science Department
Stanford University
junaraki@cs.stanford.edu

Abstract

Dialogue acts play an important role in modelling discourse phenomena in several components of modern dialogue systems. Many different features have been so far proposed for dialogue act recognition. In this report, we take a cue-based model approach, and use N-grams in utterances in dialogue as cue phrases. In our experiment with the switchboard corpus, we obtained 57.1% classification accuracy. We show that our approach is a useful technique to help us detect promising cue phrases for dialogue act recognition.

1 Introduction

Dialogue acts play an important role in modelling discourse phenomena in several components of modern dialogue systems, such as Dialogue Manager(DM)(Keizer et al., 2008), Automatic Speech Recognition(ASR)(Stolcke et al., 2000), and Text-to-Speech synthesis(TTS)(Zovato and Romportl, 2008). A dialogue act is in general taken to be composed of a dialogue act type and a semantic content. This indicates that dialogue act recognition can be formulated as a classification task of recognizing the dialogue act type given speaker's utterance(Keizer, 2003).

2 Related Work

We investigated related work in terms of two aspects: dialogue act tagsets and the cue-based model for dialogue act recognition. In this section, we mention these related work respectively.

2.1 Dialogue Act Tagsets

First, we show a list of dialogue act tagsets used or referenced in recent research in Table 1. Many research on spoken dialogue systems have used DAMSL(Allen and Core, 1997) or SWBD-DAMSL(Jurafsky et al., 1997) because of its comprehensiveness. However, some researchers pay attention to some aspects of these tagsets, and try to make some improvements to them.

One of the aspects is dimensionality of a tagset. The annotation schemes at an early stage were intended for one-dimensional annotation (exactly one tag per utterance). However, recent research argued that multidimensional tagsets (one or more tags per utterance) help to explain why utterances may have multiple functions, and are more manageable and adaptable(Petukhova and Bunt, 2009a). DAMSL was designed for multidimensional annotation, but in fact it was rarely used in such a way because many tags are supposed to be mutually exclusive. Relatively new tagsets created from such insights are DIT++(Petukhova and Bunt, 2009b) and MALTUS(Clark and Popescu-Belis, 2004).

Besides such theoretical aspects of dialogue act taxonomies, another thing to consider is whether or not a dialogue corpus associated with those taxonomies are available. Many dialogue corpora have been developed with scenario-based meetings or task-oriented conversations. Thus, utterances in these corpora can be restricted in some way. In that sense, the switchboard corpus(Jurafsky et al., 1997) gives utterances with more flexibility because they are from a set of telephone conversations on various topics without any given scenarios or tasks.

Table 1: A list of dialogue act tagsets.

Dialogue act tagsets	# of tags	Remark
AMI	16	
DIT++	86	Divided into 3 major groups.
MALTUS	13	Derived from ICSI MRDA.
ICSI MRDA	52	11 general tags and 39 specific ones.
DATE	10	
SWBD-DAMSL	42	Divided into 4 major groups.
DAMSL	32	Divided into 4 major groups.

2.2 The Cue-based Model

It is an interesting topic to consider which features are useful for dialogue act recognition. As an approach to this problem, two models have been mainly developed: the plan inference model and the cue-based model(Jurafsky and Martin, 2000). We do not explore the former model, and focus on the latter in this section.

The cue-based model is an alternative to the plan inference model, but much more attractive from a computational point of view(Keizer, 2003). Features for dialogue act classification fall into mainly the following three groups: prosodic information, words and word grammar, and discourse grammar. (Shriberg et al., 1998) examined the switchboard corpus and indicated some prosodic features such as F0 could aid dialogue act recognition. (Hirschberg and Litman, 1993) showed that certain cue words and phrases can serve as explicit indicators of discourse structure. In addition, (Kita et al., 1996) reported the effectiveness of discourse-level Hidden Markov Model (HMM) in extracting dialogue structure.

3 Approach

An overview of our approach is shown in Figure 1. As shown in this figure, our approach has two phases: the feature selection phase and the classification one. We first explain the corpus that we used in our project in the former phase in Section 3.1. We then describe each of the phases in Section 3.2 and Section 3.3, respectively.

3.1 The Corpus

We use the switchboard corpus with the SWBD-DAMSL tagset. The corpus consists of 1,155 5-minute telephone conversations. The tagset has 42 different dialogue types. With respect to features for dialogue act recognition, we focus only on N-grams of words as cue phrases, and do not consider dialogue-level sequences.

3.2 Feature Selection

Since we focus only on N-grams in an utterance, we first concatenate divided utterances into one. The corpus has some intervening utterances in a conversation as shown in Figure 2, and in this example we obtain an utterance "they almost take all emotions out of it when they report it" with a dialogue type *sv*.

We then extract all unigrams, bigrams and trigrams as cue phrases from concatenated utterances. For feature selection, we calculate pointwise mutual information (PMI) between each dialogue type and each cue phrase. Let d and c denote a dialogue type and a cue phrase. We can calculate $PMI(d, c)$ as follows:

$$PMI(d, c) = \log_2 \frac{P(d, c)}{P(d)P(c)} \quad (1)$$

In this equation, $P(d)$ and $P(c)$ are probabilities showing how often a particular dialogue type or a particular cue phrase occurs. And $P(d, c)$ is a probability for how often a particular combination of a dialogue and a cue phrase occurs simultaneously. We are interested in cue phrases with high PMI for each dialogue type, and thus select top 100 cue phrases as features.

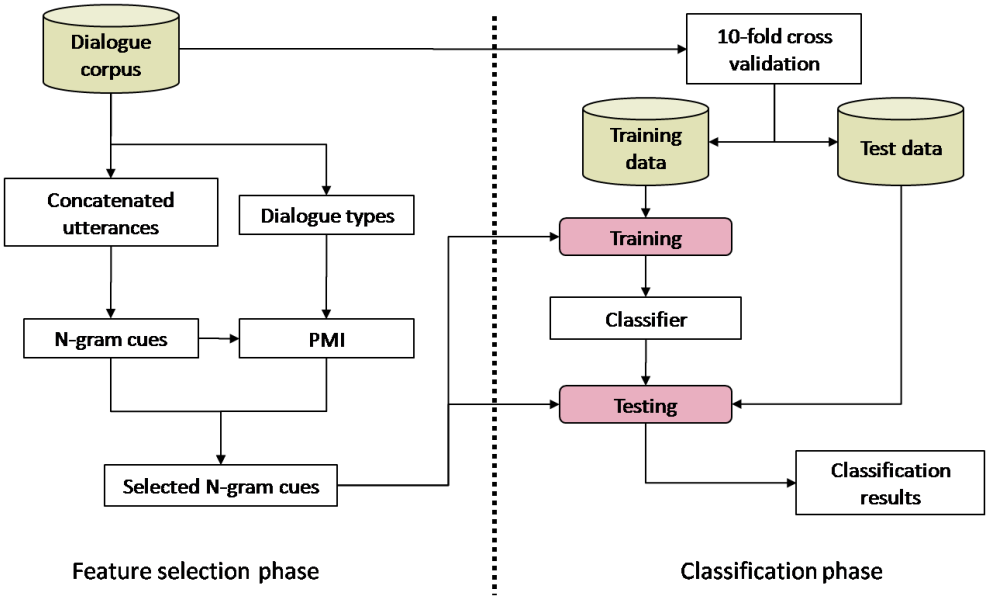

Figure 1: The feature selection and classification precess.

```
sv    A.16 utt3: they almost take all emotions out of it when they --

b^r   B.17 utt1:  Uh-huh. /

+     A.18 utt1: -- report it  /
```

Figure 2: An example of an divided utterance.

3.3 Dialogue Act Classification

We use a machine learning tool Weka(Hall et al., 2009) for applying the Naive Bayes algorithm and the multinomial logistic regression to dialogue act classification. More precisely, we use the class `weka.classifiers.bayes.NaiveBayes` for the Naive Bayes algorithm and the class `weka.classifiers.functions.Logistic` for the multinomial logistic regression. In classification, we conduct 10-fold cross validation for obtaining more reliable classification results.

4 Results

In this section, we describe what we obtained in our experiments: feature selection (Section 4.1) and dialogue act classification (Section 4.2).

4.1 Feature Selection

As described in Section 3.2, we first extracted all unigrams, bigrams, and trigrams from concatenated utterances. We show the number of all those N-grams in Table 2.

Table 2: Basic information on the switchboard corpus.

# of dialogue types	42
# of conversations	1,155
# of concatenated utterances	195,003
# of unigrams	42,350
# of bigrams	323,277
# of trigrams	720,232

We show examples of selected cue phrases and the highest PMI for some major dialogue types in Table 3. In these cue phrase examples, $\langle S \rangle$ stands for the beginning of an utterance, and $\langle /S \rangle$ for its end. We observed that most of cue phrases with high PMI are not unigrams but bigrams and trigrams. The cue phrases examples in Table 3 are relatively suitable to our conversational intuition, but actually we also observed a range of cue phrases that do not make intuitive sense. In particular, it was difficult to find out universal cue phrase in some dialogue types such as Declarative Yes-No-Question.

4.2 Dialogue Act Classification

We observed that frequencies of the cue phrases in Table 3 are very low, and every each cue phrase did not contribute greatly to classification accuracy by itself. Accordingly, instead of taking these cue phrases, we manually created possible features associated with the results, and applied them to dialogue act classification.

We show our experimental results of dialogue act classification in Table 4. We observed that the MaxEnt classifier showed almost the same performance as the Naive Bays classifier did throughout this experiment. Thus, we report only the performance of the MaxEnt classifier in that table. We accumulated possible features from ID1 to ID17, and measured classification accuracy. Thus, the accuracy increase rate in the right column shows the difference in classification accuracy between the current feature set and the previous one. As the result of this feature engineering, we obtained classification accuracy of 57.1%.

The reason why a feature dealing with a short utterance" is good is that it characterizes lots of utterances associated with dialogue types such as "Yes answers", "Other answers", "Conventional-opening", and so forth.

5 Conclusion

In this report, we focused on N-grams of words in conversations as cue phrases for dialogue act recognition. We used the switchboard corpus, and extracted various possible cue phrases for each dialogue type. As a result, we obtained 57.1% as classification accuracy in dialogue act recognition. From this experimental result, we argue that our approach with feature selection and feature engineering based on cue phrases is one of useful techniques to help us detect effective cue phrases in an efficient way. However, we need future work for refining our dialogue act recognition process in order to make the technique more useful.

6 Future Work

In this section, we make several suggestions about future work. A desirable work that readily comes to our mind is to consider some other feature selection methods besides the PMI that we used in our

Table 3: Examples of selected cue phrases and the highest PMI for each dialogue type.

Dialogue act type	Examples of selected cue phrases	The highest PMI
Statement-non-opinion	[*I just enjoy*] [*never seen any*] [*we just didn't*]	0.990
Acknowledge	[$\langle Laughter \rangle$ *Uh-huh*] [$\langle S \rangle$ $\langle Static \rangle$. *Yeah.*] [$\langle S \rangle$ $\langle Static \rangle$. *Uh-huh.*]	1.641
Statement-opinion	[*I think living*] [*believe the Social*] [*sounds like everybody's*]	2.025
Yes-No-Question	[*Did you put*] [*Do you take*] [*United States?* $\langle / S \rangle$]	3.800
Yes answers	[*Yes actually*] [*Unfortunately yes.*] [$\langle S \rangle$ *uh, yeah.*]	4.179
Wh-Question	[*Why do we*] [*What other topics*] [$\langle S \rangle$ *In what*]	4.625
No answers	[$\langle S \rangle$ *Surprisingly, no.*] [*Absolutely not.* $\langle / S \rangle$] [$\langle S \rangle$ $\langle laughter \rangle$, *no*]	4.970
Declarative Yes-No-Question	[*waist?*] [*It's something that*] [*of the leukemia?*]	5.082
Open-Question	[*How was your*] [*about your family?*] [*about you* $\langle laughter \rangle$*?*]	5.735

Table 4: Features and the resulting accuracy.

ID	Feature	Accuracy	Accuracy increase rate [%]
0	Baseline	36.8750	-
1	Short utterance ($\|u\| \leq 3$) *1	53.6801	45.57
2	Long utterance ($\|u\| \geq 10$)	53.6801	0.00
3	Ends with a question mark	55.2391	2.90
4	Ends with a exclamation mark	55.2391	0.00
5	Starts with "yes"	55.2094	-0.05
6	Starts with "no"	55.6571	0.81
7	Starts with "yeah"	55.6638	0.01
8	Contains "yes"	55.6879	0.04
9	Contains "no"	55.7509	0.11
10	Contains "yeah"	55.9438	0.35
11	Starts with "do you"	56.0166	0.13
12	Starts with "did you"	56.3197	0.54
13	Starts with 5W1H *2	56.3197	0.00
14	Contains "i think"	57.0659	1.32
15	Starts with "right"	57.0684	0.00
16	Starts with "okay"	57.0192	-0.09
17	Starts with "uh-huh"	57.0284	0.02

*1: $|u|$ stands for length of an utterance.
*2: 5W1H stands for what, where, when, why, who, and how.

project. For instance, it might be good to consider a probability such as $P(d|c)$.

The next thing that we can do is to consider syntactic information for an utterance given by some syntactic parser. For example, utterances starting with "Can you" or "Would you" are likely to show requests. Considering modal verbs at the head as features might be helpful to increasing the classification accuracy.

Another things to do are to consider other features from prosodic information and discourse grammar for more disambiguation for features. For instance, a sequence of a question and an answer is probably a good factor to be extracted.

References

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers. Technical report, Multiparty Discourse Group. University of Rochester.

Alexander Clark and Andrei Popescu-Belis. 2004. Multi-level Dialogue Act Tags. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 163–170. Association for Computational Linguistics.

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

Julia Hirschberg and Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall.

Daniel Jurafsky, Liz Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13. Technical report, University of Colorado at Boulder Technical Report 97-02.

Simon Keizer, Milica Gasic, Francois Mairesse, Blaise Thomson, Kai Yu, and Steve Young. 2008. Modelling user behaviour in the HIS-POMDP dialogue manager. In *IEEE SLT*, pages 121–124, December.

Simon Keizer. 2003. *Reasoning under uncertainty in natural language dialogue using bayesian networks*. Dissertation, Twente University.

Kenji Kita, Yoshikazu Fukui, Masaaki Nagata, and Tsuyoshi Morimoto. 1996. Automatic acquisition of probabilistic dialogue models. In *ICSLP-96*, volume 1, pages 196–199.

Volha Petukhova and Harry Bunt. 2009a. The independence of dimensions in multidimensional dialogue act annotation. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 197–200. Association for Computational Linguistics.

Volha Petukhova and Harry Bunt. 2009b. Towards a Multidimensional Semantics of Discourse Markers in Spoken Dialogue. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 157–168. Association for Computational Linguistics, January.

Elizabeth Shriberg, Rebecca Bates, Paul Taylor, Andreas Stolcke, Daniel Jurafsky, Klaus Ries, Noah Coccaro, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. 1998. Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? *Language and Speech*, 41(3-4):439–487.

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*, 26:339–373.

Enrico Zovato and Jan Romportl. 2008. Speech synthesis and emotions: a compromise between flexibility and believability. In *Proceedings of Fourth International Workshop on Human-Computer Conversation*.