"Dialog Systems" class, Spring 2002

- TERM PROJECT -

Dialog Act Tagging using Memory-Based Learning

Mihai Rotaru
Computer Science Department, University of Pittsburgh,
mrotaru@cs.pitt.edu

Abstract

We are applying a memory based learning (MBL) algorithm to the task of automatic dialog act (DA) tagging. This work is along the lines of a recent trend that considers MBL as being more appropriate for natural language processing. We did the experiments on the Switchboard corpus, overcome the problem of feature selection and yield results that seem to be better that previous reported results on the same corpus.

1. Introduction

A better understanding of the semantics of user utterances in a spoken dialog system (SDS) will lead to more efficient and robust SDS. Given the fact that deep semantic analysis of sentences still eludes us, researchers have turned their eyes to a simpler problem: shallow semantics analysis. Dialog act (DA) tagging is one type of shallow semantics analysis of sentences. DAs are a concise abstraction of utterance semantics and usually express the illocutionary force of the utterance (SUGGEST, ACCEPT, QUESTION etc.).

Different DA schemes have been devised in recent years, some relevant to particular problems, other aiming for domain-independence. Notable is the DAMSL architecture created by Discourse Resource Initiative (Core and Allen 1997). This architecture was extended and used to tag a human-human conversation corpus, Switchboard. The resulting DA scheme is known as SWBD-DAMSL and contains approximately 220 different DAs. DAs were obtained by combining basic DAs with orthogonal functions of the utterance (like task-management related, communication-management related). Due to low frequency of some DAs, the tag set was further clustered to a set of 42 relevant DAs (Jurafsky et all, 1997). We used this dataset along with clustered SWBD-DAMSL architecture in our experiments.

Even if it does a shallow understanding of the utterances, DA tagging is useful in many applications: detection dialog game boundaries, detecting the interaction dominance and discussion genre. Also, a meeting summarizer should be able to detect who spokes, to whom is the conversation addressed and if it asked a question or answered one. Even more, DAs may help the speech recognizer, by restricting its grammar to the one corresponding to the expected DA.

One disadvantage of the DAs is the fact that inter-coder reliability measured by kappa-statistic (Carletta 1996) is usually not as high as expected. This phenomenon can be caused by the shallowness nature of the DA and the fact that people give different abstraction of semantics for different utterances if the available spectrum of DAs is small or inappropriate. This should also extend to the accuracy of an automatic DA tagger.

The rest of the project paper is organized as follows. In Section 2, we discuss other approaches to automatic DA tagging. Section 3 states the reasons for using MBL and we give a brief description of MBL techniques used and their advantages. In section 4 we described our processing of Switchboard corpus, the way we selected features for the MBL learner and the experiment setup. In section 5 we present the results we obtained. The last section draws the project conclusion and addresses future work.

2. Previous work on automatic DA tagging

Given the potential advantages of DAs, many researchers approached the task of automatic DA tagging: an algorithm capable of labeling user utterances with DAs given a set of evidences (like ASR output, prosodic information, dialog manager state, etc.).

One direction in this research is driven by the observation that utterance DAs are correlated with word substrings (words or small word phrases) that appear in the utterance. These word substrings are called cue-phrases and have been identified in work done by Hirshberg and Litman (1993). One problem of cue-phrases is that they are usually domain and task dependent. To overcome these problems, Reithinger and Klesen (1997) used word n-grams (all word substrings with n words) obtained from the corpus. Their approach had the problem that a large number of feature was devised by word n-grams and the machine learning algorithm had to be able to discard irrelevant features. As a happy medium between above approaches is the work of Samuel et al (1998, 2000) that uses cue-phrases and a subset of the word n-grams called dialog act cues. Combined with clustering techniques for dialog act cues, their results peaked at 77.44% accuracy for the VERBMOBIL corpus.

Another direction in automatic DA tagging is the statistical approach. Outstanding is the work of Shriberg et all (1998) and Stolcke et all (2000) that modeled the dialog structure of conversation as an HMM (Hidden Markov Model). DAs were the hidden state while user sentences (ASR output) were the observation. The advantage of this approach is that it can also incorporate other dialog information (like prosodic) which can further increase the accuracy. Their best result on the Switchboard corpus was a accuracy of 71% with the SWBD-DAMSL architecture (Shriberg et all, 1998).

3. Memory-Based Learning (MBL)

MBL is a direct application of memory-based reasoning theory in the area of machine learning. The theory is based on the hypothesis that people handle a new situation by matching them with stored representation of previous situations. The matching process assumes a similarity metric between situations. This theory is opposite to the one that attributes people behavior to the application of mental rules abstracted from previous situations. Since no processing is done on previous experiences, these types of learning are called lazy learning.

An advantage of MBL is that in can handle very well domains that present exceptions and sub-regularities (usually considered noise by other machine learning

techniques). It was proven that natural language processing is such type of domain by the work done by Daelemans et all (1999) and Van den Bosch (2001). They showed that keeping exceptional instances in memory is the reason for better accuracy of MBL over other machine learning techniques that prune them (like rule-based learner or decision trees).

The disadvantage of MBL is its high memory and computational demands. Memory problem comes from its advantage: all instances are kept in memory for future reference. The computational problem comes from algorithm 'laziness': a new instance has to be compared (with respect to similarity) with all stored instances, which for a big training set can become a serious problem. Smart indexing of training instances can alleviate these problems.

Based on the proved advantages of MBL in natural language processing task, our idea was to test if these advantages are still valid for the task of automatic DA tagging. For our experiments, we employed the IB1-IG implementation of MBL from TiMBL software package (Daelemans et all 2001). This a k-NN algorithm implementation with weights computed based on gain ratio automatically computed from train test.

4. Switchboard corpus and feature extraction for MBL

The original idea came while reading Samuel et all (1998) paper. Their experiments where run on VERBMOBIL corpus. Due to lack of access to the VERBMOBIL tagged transcripts we had to switch to a publicly available corpus, the Switchboard corpus.

We run the experiments on a dataset of more that 1200 hand labeled dialogs from the Switchboard corpus of spontaneous human-human telephone speech. Every utterance from the corpus was labeled with respect to its DA using SWBD-DAMSL architecture. Best results on this corpus were those reported by Stolcke et all (2000): a 71% accuracy with a trigram word model.

Our intention was to challenge these results with the MBL approach. Since former experiments were done on a clustered set of 42 DAs and available dataset had all 220 DA, we had to do the DA clustering as described in Jurafsky et all (1997). Clustering the DAs resulted in a total number of 43 DA: the regular 42 DAs plus "+" DA. The "+" DA was used as an equivalent DA for DAMSL "Segment". The tag had a high frequency (more than 8%) and was ignored in Stolcke's experiment. We choose to do our experiments both with this tag and without it.

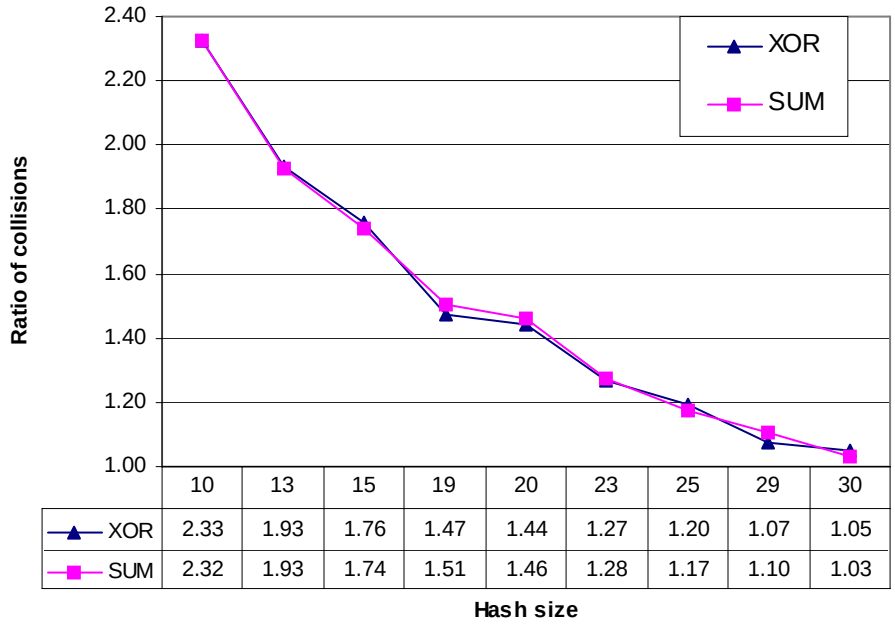
The resulting data set had more than 210,000 utterances if "+" DA was used and 200,000 if DA "+" was not used. We did the experiments by cutting out a test data of 5000 utterances from the dataset (compared with 4000 from Stolcke's study). It is very important to notice that the two experiments cannot be compared *directly*. This is due to a different train/test set. Even more, both show a trend in the ability of the underlying models because no cross-validation was done.

The biggest problem we faced was devising a relevant set of features. From the original corpus it was easy to extract features like: the dialog acts of previous three utterances and the number of utterances since last change of speaker. But these features can only model the sequencing of DAs. A shallow semantic of the utterance needs to be modeled also. In Stolcke's experiment this was modeled by class conditional probability of word n-grams. One idea will be to use a set of features that tell if a DA cue is present or not in the utterance. Given the high number of DA cues that have to be considered and the huge size of dataset, this approach is infeasible.

Our idea was to represent word bigrams as features. The intuition behind this is the fact that when comparing two instances (utterances) the number of common bigram should be correlated with the equality of utterances DAs. Even more, given the size of the train set, for any new utterance we can hope that there exists in the train set an instance with common bigrams and the same DAs (this is based on the assumption that DAs are determined by specific lexical constructs).

The next problem was TiMBL inability to handle set features. Out set feature would have been the set of bigrams from the utterance. Setting a feature with a bigram was not a solution because there is no inter-feature comparison in TiMBL (in order to count the number of common bigrams). To overcome this problem we had to use the following trick: we hashed all bigrams to a given number of features (slots). Hashing was done using a function of the letters present in the bigrams and the number of hash slots. Hash collisions were "resolved" by replacing the old value. This ensured that any bigram would be present in the same slot (feature) for any utterance that contained the bigram.

We used several values for the hash size and hash function. We experimented with hash size of 10,13,15,19,20,23,25,29 and 30. Hash functions available were either a sum or a bit-wise exclusive OR of all character codes present in bigram. Figure 1, plots the ratio of hash collisions per utterance when varying hash parameters. We choose to do experiments on 3 versions: 13 and 30 hash size with SUM hash function and 20-hash size with XOR hash function.



	10	13	15	19	20	23	25	29	30
XOR	2.33	1.93	1.76	1.47	1.44	1.27	1.20	1.07	1.05
SUM	2.32	1.93	1.74	1.51	1.46	1.28	1.17	1.10	1.03

Figure 1 – Ratio of hash collisions per utterance. No significant differences between the two hash function. As expected, ratio decreases with hash size.

The last dimension to take in consideration in our experiments was the value to be set for empty hash slots (features). There were two possibilities: use the same value or different values for train and test data empty slots. The semantic difference is big: if same values are used, corresponding empty slots from train and test have the same impact as common bigrams. The other version does not have this anomaly and this will reflect on our results too.

5. Results

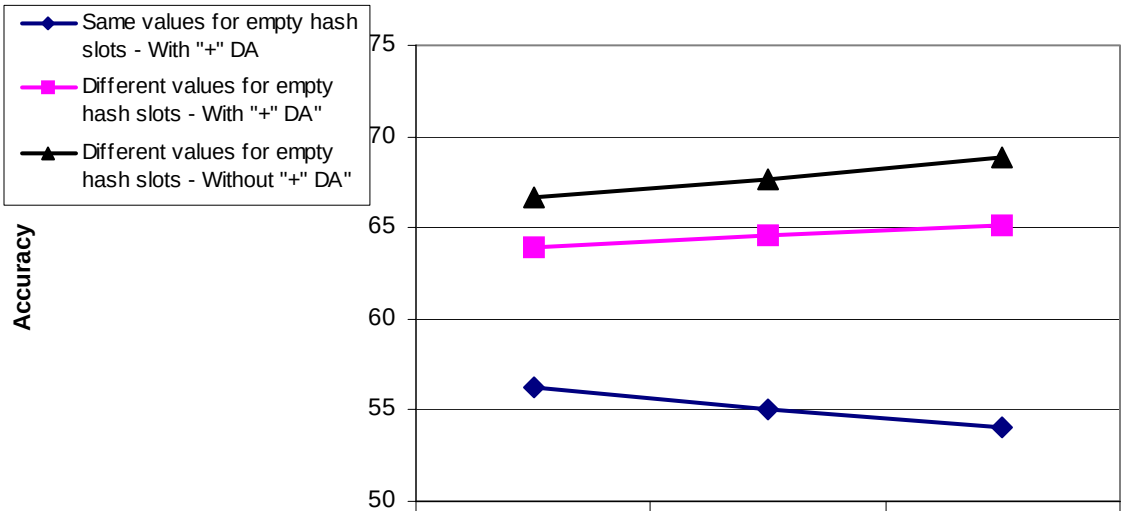
There are several dimensions we analyzed in our experiments:

- Hash structure (3 possibilities),
- Values for empty hash slots and
- Presence of "+" DA.

We compared the experiments with a majority baseline of 35% (always selecting *sd* DA) and previous result on the same corpus of 71% accuracy obtained in Stolcke's xperiments.

First experiments we did compare the influence of first two dimensions ("+" DA was used). Figure 2 detail results. We can see that choosing the same value for empty slots decreases the performance. This is explicable based on the same value anomaly (see Section 5). An increase in hash size is followed by a slight increase in accuracy, but the increase is not compensated by the overwhelming increase in memory needed.

We also experimented the results with "+" DA present or not (i.e "+" labeled sentences were pruned out or not). The accuracy without "+" DA is better that the case with "+" DA present and the hash size has a bigger influence on the accuracy.



	13	20	30
Same values for empty hash slots - With "+" DA	56.3	55.06	54.1
Different values for empty hash slots - With "+" DA"	63.94	64.56	65.16
Different values for empty hash slots - Without "+" DA"	66.64	67.6	68.84

Figure 2. Accuracy of MBL based on hash size, values for empty hash slots and usage of "+" tag.

We choose to test also the influence of the number of neighbors used in MBL (all previous experiments where with one neighbor). For this experiment we used the hash with size 30 and without "+" DA dataset. Figure 3 shows the results. Our highest performance was 72.32% accuracy with three neighbors, which is better

than Stolcke's results. Further increase in number of neighbors did not improve the accuracy.

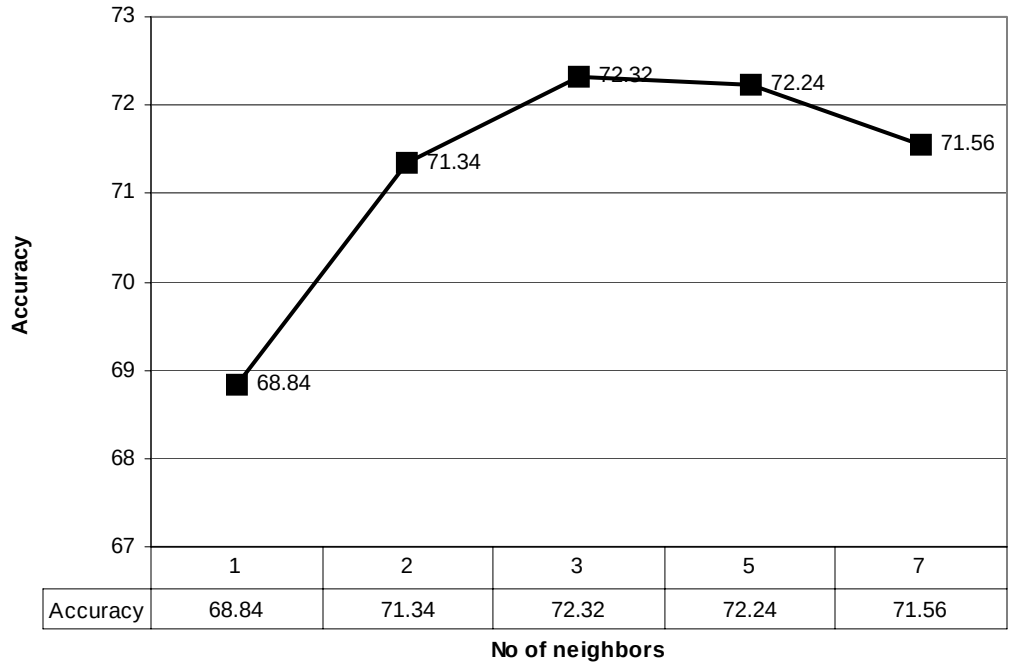


Figure 3. The effect on accuracy of the number of neighbors used.

A confusion matrix for our best performer (3 neighbor, hash size 30, without "+" DA) is presented in Table1. A lot of accuracy is lost when trying to distinguish between statements with no opinion (*sd* DA) and statements with opinion (*sv* DA). Recognizing backchannel DA (*ba*) is more accurate especially due to limited forms of backchanneling (usually those utterances contain only "Uh-huh"), also there is a great deal of confusion between it and Accept/Agree DA (aa).

	sd	**sv**	**b**	**%**	**aa**
sd	1667	154	4	42	9
sv	370	327	1	36	13
b	1	0	809	6	18
%	58	19	21	223	3
aa	11	12	99	3	132

Table1. Confusion matrix for high frequency classes (DAs)

We continued experiment on the best performer by using only DA history features or only bigrams. The first one yield an accuracy of only 46.48% while the second one 63.72%. This proves the importance of bigram features for our task.

6. Conclusions

We showed that the advantage of MBL on natural language processing seems to extend also the task of automatic dialog act tagging. Further studies are required to confirm the hypothesis and deep understanding of why MBL works better should be pursued.

Future work includes trying to explain why three neighbors yields higher performance, modifying TiMBL to support set features (in this way no bigrams will be missed and there will be a big decrease in memory usage).

I would like to thank to Ken Samuel (for providing me with dialog cues) and Diane Litman. Special appreciation for to my wife Diana for bearing with me working all day long.

References

Antal van den Bosch, Emiel Krahmer, and Marc Swerts. "Detecting problematic turns in human-machine interactions: Rule-induction versus memory-based learning approaches" In Proceedings of the 39th Meeting of the Association for Computational Linguistics (ACL'00). New Brunswick, NJ: ACL, pp. 499-506, 2001.

Core, Mark and James Allen "Coding Dialogs with the DAMSL Annotation Scheme." AAAI Fall Symposium on Communicative Action in Humans and Machines, 1997.

Daelemans Walter, Antal van den Bosch, and Jakub Zavrel. "Forgetting exceptions is harmful in language learning" Machine Learning, special issue on natural language learning, 34, pp. 11-43, 1999.

Daelemans Walter, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch "TiMBL: Tilburg Memory Based Learner, version 4.0, Reference Guide" Reference: ILK Technical Report 01-04, 2001.

Hirshberg, Julia and Litman, Diane "Empirical studies on the disambiguation of cue phrases". Computational Linguistic 12(3):175-204, 1993

Jean Carletta "Assessing Agreement on Classification Tasks: The Kappa Statistic". Computational Linguistics, 22(2):249-254, 1996.

Jurafsky, Daniel, Elizabeth Shriberg, and Debra Biasca. (1997) "Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13" (postscript) (HTML) University of Colorado, Boulder. Institute of Cognitive Science Technical Report 97-02

Reithinger, Norbert and Klesen, Martin "Dialog act classification using language models." Proceedings of 5th European Conference on Speech Communication and Technology, 1997

Samuel Ken, Sandra Carberry, and K. Vijay-Shanker." Dialogue Act Tagging with Transformation-Based Learning" Proceedings of the 36th Meeting of the Association for Computational Linguistics (ACL), 1998

Samuel, Ken. Discourse Learning: An Investigation of Dialogue Act Tagging using Transformation-Based Learning. Ph.D. Dissertation. Department of Computer and Information Sciences. University of Delaware. Newark, Delaware. 2000

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

Stolcke Andreas, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech". Computational Linguistics 26:3, 2000.