DIALOG ACT TAGGING USING GRAPHICAL MODELS

Gang Ji and Jeff Bilmes

University of Washington
Department of Electrical Engineering
Seattle, WA 98105
{gang,bilmes}@ssli.ee.washington.edu

ABSTRACT

Detecting discourse patterns such as dialog acts (DAs) is an important factor for processing spoken conversations and meetings. Different techniques have been used to tag dialog acts in the past such as hidden Markov models and neural networks. In this work, a full analysis of dialog act tagging using different generative and conditional dynamic Bayesian networks (DBNs) is performed, where both conventional switching n-grams and factored language models (FLMs) are used as DBN edge implementations. Our tests on the ICSI meeting recorder dialog act (MRDA) corpus show that the factored language model implementations are better than the switching n-gram approach. Our results also show that by using virtual evidence, the label bias problem in conditional models can be avoided. Also, we find that on a corpus such as MRDA, using the dialog acts of previous sentences to help predict current words does not improve our conditional model.

1. INTRODUCTION

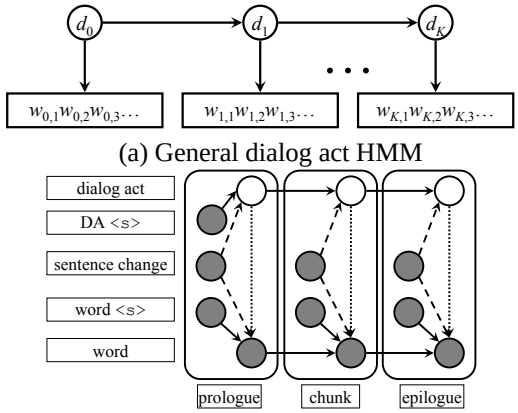
A conversation or meeting contains a particular kind of discourse structure, known as the *dialog act* (DA). Dialog acts reflect the functions that utterances serve in a discourse. An utterance can serve as a statement, question, or acknowledgment of another speaker's contributions. It has been shown that DAs are very important to problems such as speech recognition [1], spoken language understanding (SLU) [2], and dialog translation [3].

There have been many attempts to build stochastic models for dialog structure including hidden Markov models [1], neural networks [4], fuzzy fragment-class Markov models [5], semantic classification trees and polygrams [6], and maximum entropy models [7]. However, a complete analysis of the use of generative and conditional dynamic Bayesian networks (DBNs) [8] for tagging DAs is missing.

In this work, DA tagging with different types of DBNs is carried out with the help of the graphical models toolkit (GMTK) [9], a DBN system for speech, language, and time series data. We also compare both switching n-gram models and factored language models (FLMs) [10] as implementations of edges in our DBNs. Our results on the ICSI meeting recorder dialog act (MRDA) corpus [11] indicate that FLMs with an appropriate backoff path perform better than switching n-gram models in all cases. Furthermore, we analyze several conditional models and show that while simple conditional models suffer from label bias [12], this can be corrected using virtual evidence [13, 14] without leaving the directed graphical modeling paradigm.

The paper is organized as follows: Section 2 provides an analysis of generative models using the two edge implementations mentioned above. Section 3 introduces several conditional models, the first with and the remainder without label bias. Finally, Section 4 concludes.

2. GENERATIVE MODELS FOR DIALOG ACT TAGGING



(a) General dialog act HMM

(b) Specific dialog act DBN with word bigram

Fig. 1. Generative models for dialog act tagging.

Like most tagging tasks, both generative and conditional models can be used. In this section, we focus on generative

This work was funded by NSF under grant IIS-0121396.

models. Figure 1 shows a generative model (a) and its more specific form (b) that is actually used: the prologue is the first, the epilogue is the last, and the chunk is the repeated DBN frame unrolled to fit the entire observation length. The general DBN shows that sentences are conditionally independent given DAs. The specific DBN functions as follows: The binary switching parent [9] "sentence change" indicates when a new DA is chosen, and when it stays fixed between successive time intervals — when a sentence change occurs, the next DA is chosen based on a DA bi-gram, otherwise the dialog act is a copy of the previous DA. The sentence change also indicates when the word variable should ignore the previous word (from the previous sentence) and condition on the special start-of-sentence token "<s>" instead (which is also used in the prologue for a virtual first DA). In either case, the word uses a word-bigram model. In this work, we assume that sentence change is observed as is typical for this corpus. In general, this observation could come from an automatic segmentation. Alternatively, the variable could be hidden as part of a joint segmentation/DA labeling.

2.1. Edge Implementation

Graphical models specify the underlying model family, but they do not explicitly provide edge implementations. In fact, there are many ways to implement such a dependency. For example, the dialog act bigram could be implemented using a dense conditional probability table (CPT) estimated using maximum-likelihood. In this paper, we investigate more sophisticated implementations, namely switching n-grams, and factored language models.

According to Figure 1(b), the i-th word $w_{k,i}$ is modeled using $P(w_{k,i}|w_{k,i-1}, d_k)$, where d_k is the DA for sentence k (note that this is true for both values of sentence change). A standard way to implement this CPT is to train separate word bigrams $P_d(w_i|w_{i-1})$ for each value of d_k, thus effectively splitting the training data into D disjoint groups (D is the number of possible DAs). We can thus view the DA variable as a switching parent in a DBN, switching in the appropriate word-only CPT. A potential problem, however, is that some word models might end up being trained with very little data.

To mitigate this problem, FLMs and generalized backoff can be used. Here, words and dialog acts are treated as different factors in the same language model $P(w_{k,i}|w_{k,i-1}, d_k)$. There are several advantages of using this approach. First, as a unified model, it can be trained on the entire training set without needing forced data splitting, thus avoiding data sparseness if the right backoff path is used. In this work, we first drop $w_{k,i-1}$ leading to the backoff model $P(w_{k,i}|d_k)$ so that the important relation for discrimination (namely, the connection between words and DAs) is retained for as long as possible. More importantly, the smoothing method (modified Kneser-Ney [15] in this case) smoothes over all the data, rather than just over the split data in the switching case. Last, generalized parallel back-off [10] can also be used (it is not employed in this work however).

2.2. Generative Model Experiments

We tested the different approaches described above on the ICSI Meeting Recorder Dialog Act (MRDA) corpus [11]. The data contains 75 naturally recorded meeting conversations on different topics. It has 116,555 sentences and 759,837 word tokens in total. The vocabulary size is 14,347 and there are 1,260 unique dialog acts in the corpus. Each dialog act contains a main tag, several optional special tags, and an optional disruption form. Because the number of unique dialog acts is too large given the available data, we removed the special tags thereby reducing the number of unique dialog acts to 62, a reasonable size for this task. In our experiments, 65 randomly selected meetings were used for training and the remaining 10 meetings were used for testing. Even with the reduced DA set size, there were still 13 DAs with only one sentence to train on, and 5 DAs with only two sentences to train on.

In addition to the word bigram models shown in our graph, we also tried a word unigram and a word trigram (both switching and FLM). In all cases a DA bigram is used. All language models were trained by SRILM [16] and the FLM SRILM extensions [10]. The graphical models toolkit (GMTK) [9][1] was used to test all models. The results are presented in Table 1

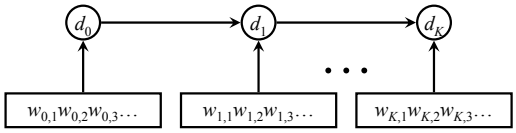
Table 1. Comparison between switching n-grams and FLMs

model	word	accuracy
switching n-gram	unigram	63.6%
	bigram	64.2%
	trigram	46.2%
FLM	unigram	64.1%
	bigram	65.0%
	trigram	66.0%

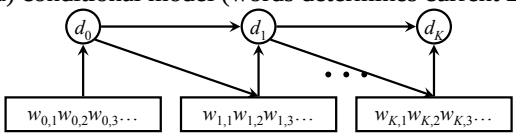
The FLM is consistently better than the switching n-gram. In fact, in the switching n-gram case, the word trigram is worse than the bigram because after splitting, each DA specific trigram has only a tiny amount of training data. In fact, since some dialog acts contain only one or two training sentences, the resulting word language models will be highly skewed. With the FLM, however, language model estimation is much more robust. Unlike the switching n-gram, the trigram FLM has the best overall accuracy.

[1]The version we used is a completely re-written version, and that also now supports native mode standard ARPA language model and FLMs as CPT implementations.

3. CONDITIONAL MODELS AND LABEL BIAS



(a) conditional model (words determines current DA)



(b) conditional model with help of previous DA

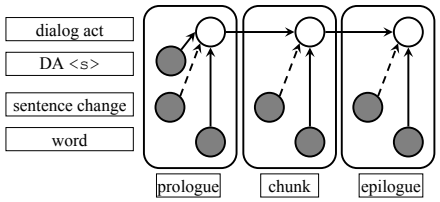
Fig. 2. Models for dialog act tagging.

In addition to the generative models in the previous section, conditional models [12] can also been used, and Figure 2 shows two general possibilities. In Figure 2(a), the DA is decided by the words in the current sentence in addition to the previous dialog act. In this case, the words in one sentence will be independent of the dialog acts of all previous sentences, not an assumption made by the generative models in Figure 1. This, in fact, can lead to the so-called label bias problem [12]. We can remove this problem by adding a link from the DA of a previous sentence to the current sentence as shown in Figure 2(b).

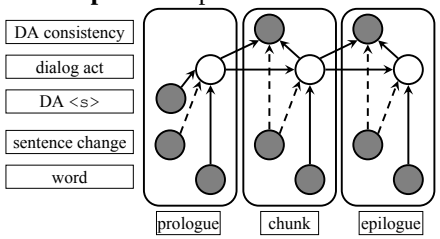
In fact, there are several ways to remove label bias while retaining the use of Bayesian networks. Figure 3 shows the DBNs we actually used in our experiments. Model *a-simple* is a simple conditional model exhibiting label bias, where the dialog acts are predicted by the current word and the previous dialog act. Label bias is particularly severe here since when there is no sentence change, the DA will keep its previous value and ignore the current word entirely. This means that the DA choice is unaffected by other than the first word of each sentence.

We propose a novel way to fix this problem using virtual evidence [13, 14] as shown in *a-virtual*. Here, when a sentence change occurs, the new dialog act will be predicted by the current word and previous dialog act. Otherwise, it will be predicted by the current word only. We use, however, the virtual evidence variable "DA consistency" in this case to guarantee this DA is equal to its previous value. Therefore, the score of the DA given the current word at each frame will be accumulated together all through the sentence. But because words are no longer independent of the dialog act of the previous frame, the entire accumulated sentence score (rather than 1 word) decides the DA, and hence label bias is avoided.

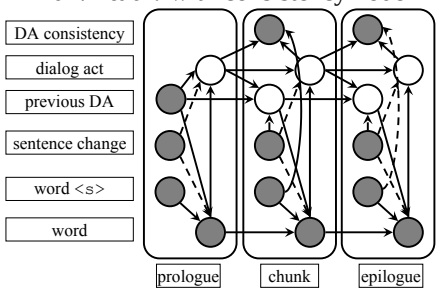
As can be seen, in both models *a-simple* and *a-virtual*, there is no between-word edge. This is because such links will have no effect on our decoding result since their score is identical for all dialog act hypotheses. In model *b-virtual*, we add the use of the previous DA for predicting the current sentence (like the general case in Figure 2(b)). In this case, between-word edges could possibly have an effect. In our experiments, only the word bigram is used.



a-simple: a simple conditional model



a-virtual: with consistency node



b-virtual: with previous DA and consistency node

Fig. 3. Graphical models for dialog act tagging with the graphical model toolkit.

Table 2. Comparison among different conditional models

model	accuracy
a-simple	34.5%
a-virtual	64.1%
b-virtual	63.6%

Like before, these graphical models were tested on the MRDA corpus with the same experimental setup described before. In Table 2, our experimental results are presented. As expected, the simple conditional model *a-simple* gives us poor accuracy due to severe label bias. When adding a virtual evidence random variable in model *a-virtual*, we get an enormous improvement over *a-simple*. Lastly, we see a slight decrease in performance using *(b-virtual)*. In *(b-virtual)* (and Figure 2(b)), words in one sentence are not independent of the previous sentence's dialog act, and therefore one might expect to see improvement. Our results show,

however, that the use of the previous DA apparently hurts performance. Our belief is that training data sparseness is the culprit — using FLMs and advanced smoothing helps but not enough to compensate for such a small relative training data size for this model.

Comparing these results with those of the previous section, the generative model with a trigram FLM is best. The unigram FLM (with an accuracy of 64.1%) matches *a-virtual*'s performance. Our contention is that the generative model's ability to represent DA-specific word sequence information (via the bigram and trigram FLM) helps significantly in increasing DA tagging accuracy.

4. DISCUSSION

We have presented a variety of DBNs for the dialog act tagging task on the ICSI meeting recorder dialog act (MRDA) corpus. In our generative model, it has been shown that a factored language model (FLM) implementation with an appropriate backoff path out-performs switching n-grams. In our conditional model, results show that by adding a virtual evidence random variable, one can avoid the label bias problem. In MRDA, adding a link from the previous sentence's DA to the words of the current sentence does not increase performance, presumably due to training data sparseness. Among all the models, our generative model with a word trigram FLM has the best accuracy, and beats the conditional model because it does a better job modeling the DA specific word sequences. In future work, we will extend the conditional model with an ability to represent similar such word sequence information.

In our experiments, language model penalties were not used. The probabilistic models for words and dialog acts also have the same weights. In future work, we will tune penalties and weights on a development set in an attempt to further improve accuracy. Moreover, we will apply DA tagging on a bigger corpus, such as Switchboard [17, 1]. With more data, we will discover if using the link from the previous dialog act to words in the current utterance will indeed increase the performance of our conditional model, as we expect. With the power of GMTK, it is easy to quickly try many different models with only the addition of an edge or two. In addition, it will also be informative to compare the above models with conditional random fields [18, 7].

The authors would also like to thank Chris Bartels and Simon King for their GMTK graph triangulation scripts, and Sheila Reynolds and Jon Malkin for useful comments.

5. REFERENCES

[1] A. Stolcke *et al.*, "Dialog act modeling for conversational speech," in *Proc. of the AAAI Spring Symp. on Appl. Machine Learning to Discourse Processing*, 1998, pp. 98–105.

[2] Y. He and S. Young, "A data-driven spoken language understanding system," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003, pp. 583–588.

[3] H. Lee, J.-W. Lee, and J. Seo, "Speech act analysis model of Korean utterances for automatic dialog translation," *Journal of KISS(B) (Software and Applications)*, vol. 25, no. 10, pp. 1443–1452, 1998.

[4] H.-F. Wang, W. Gao, and S. Li, "Dialog act analysis of spoken Chinese based on neural networks," *Chinese Journal of Computers*, vol. 22, no. 10, pp. 1014–1018, 1999.

[5] C.-H. Wu, G.-L. Yan, and C.-L. Lin, "Speech act modeling in a spoken dialog system using a fuzzy fragment-class Markov model," *Speech Communication*, vol. 38, no. 1-2, pp. 183–199, 2002.

[6] M. Mast *et al.*, "Automatic classification of dialog acts with semantic classification trees and polygrams," *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pp. 217–229, 1996.

[7] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. of ICASSP*, 2005.

[8] K. Murphy, *Dynamic Bayesian Networks, Representation, Inference, and Learning*, Ph.D. thesis, MIT, Department of Computer Science, 2002.

[9] J. Bilmes and G. Zweig, "The graphical models toolkit: An open source software system for speech and time-series processing," in *Proc. of ICASSP*, 2002, vol. 4, pp. 3916–3919.

[10] J. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel backoff," in *Proc. of HLT-NACACL: Short Papers*, 2003, pp. 4–6.

[11] E. Shriberg *et al.*, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. of the 5th SIGdial Workshop on Discourse and Dialogue*, 2004, pp. 97–100.

[12] D. Klein and C. Manning, "Conditional structure versus conditional estimation in NLP models," in *Proc. Empirical Methods in Natural Language Processing*, 2002, pp. 9–16.

[13] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 2nd printing edition, 1988.

[14] J. Bilmes, "On soft evidence in bayesian networks," Tech. Rep. UWEETR-2004-0016, U. Washington Dept.of Electrical Engineering, 2004.

[15] Stanley F. Chen and Joshua Goodman, "An empirical study of smoothing techniques for language modeling," Tech. Rep. TR-10-98, Computer Science Group, Harvard University, August 1998.

[16] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. ICSLP*, 2002, vol. 2, pp. 901–904.

[17] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. of ICASSP*, 1992, pp. 517–520.

[18] C. Sutton, K. Rohanimanesh, and A. McCallum, "Dynamic conditional random fields: factorized probabilistic models for labeling and segmenting sequence data," in *Proc. of ICML*, 2004.