Decision detection using hierarchical graphical models

Trung H. Bui
CSLI
Stanford University
Stanford, CA 94305, USA
thbui@stanford.edu

Stanley Peters
CSLI
Stanford University
Stanford, CA 94305, USA
peters@csli.stanford.edu

Abstract

We investigate hierarchical graphical models (HGMs) for automatically detecting decisions in multi-party discussions. Several types of dialogue act (DA) are distinguished on the basis of their roles in formulating decisions. HGMs enable us to model dependencies between observed features of discussions, decision DAs, and subdialogues that result in a decision. For the task of detecting decision regions, an HGM classifier was found to outperform non-hierarchical graphical models and support vector machines, raising the F1-score to 0.80 from 0.55.

1 Introduction

In work environments, people share information and make decisions in multi-party conversations known as meetings. The demand for systems that can automatically process information contained in audio and video recordings of meetings is growing rapidly. Our own research, and that of other contemporary projects (Janin et al., 2004) aim at meeting this demand.

We are currently investigating the automatic detection of decision discussions. Our approach involves distinguishing between different dialogue act (DA) types based on their role in the decision-making process. These DA types are called Decision Dialogue Acts (DDAs). Groups of DDAs combine to form a decision region.

Recent work (Bui et al., 2009) showed that Directed Graphical Models (DGMs) outperform other machine learning techniques such as Support Vector Machines (SVMs) for detecting individual DDAs. However, the proposed models, which were non-hierarchical, did not significantly improve identification of decision regions. This paper tests whether giving DGMs hierarchical structure (making them HGMs) can improve their performance at this task compared with non-hierarchical DGMs.

We proceed as follows. Section 2 discusses related work, and section 3 our data set and annotation scheme for decision discussions. Section 4 summarizes previous decision detection experiments using DGMs. Section 5 presents the HGM approach, and section 6 describes our HGM experiments. Finally, section 7 draws conclusions and presents ideas for future work.

2 Related work

User studies (Banerjee et al., 2005) have confirmed that meeting participants consider decisions to be one of the most important meeting outputs, and Whittaker et al. (2006) found that the development of an automatic decision detection component is critical for re-using meeting archives. With the new availability of substantial meeting corpora such as the AMI corpus (McCowan et al., 2005), recent years have seen an increasing amount of research on decision-making dialogue. This research has tackled issues such as the automatic detection of agreement and disagreement (Galley et al., 2004), and of the level of involvement of conversational participants (Gatica-Perez et al., 2005). Recent work on automatic detection of decisions has been conducted by Hsueh and Moore (2007), Fernández et al. (2008), and Bui et al. (2009).

Fernández et al. (2008) proposed an approach to modeling the structure of decision-making dialogue. These authors designed an annotation scheme that takes account of the different roles that utterances can play in the decision-making process—for example it distinguishes between DDAs that initiate a decision discussion by raising an issue, those that propose a resolution of the issue, and those that express agreement to a proposed resolution. The authors annotated a portion of the AMI corpus, and then applied what

Proceedings of the ACL 2010 Conference Short Papers, pages 307–312,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

they refer to as “hierarchical classification.” Here, one *sub-classifier* per DDA class hypothesizes occurrences of that type of DDA and then, based on these hypotheses, a *super-classifier* determines which regions of dialogue are decision discussions. All of the classifiers, (sub and super), were linear kernel binary SVMs. Results were better than those obtained with (Hsueh and Moore, 2007)’s approach—the F1-score for detecting decision discussions in manual transcripts was 0.58 *vs.* 0.50. Purver et al. (2007) had earlier detected action items with the approach Fernández et al. (2008) extended to decisions.

Bui et al. (2009) built on the promising results of (Fernández et al., 2008), by employing DGMs in place of SVMs. DGMs are attractive because they provide a natural framework for modeling sequence and dependencies between variables, including the DDAs. Bui et al. (2009) were especially interested in whether DGMs better exploit non-lexical features. Fernández et al. (2008) obtained much more value from lexical than non-lexical features (and indeed no value at all from prosodic features), but lexical features have limitations. In particular, they can be domain specific, increase the size of the feature space dramatically, and deteriorate more in quality than other features when automatic speech recognition (ASR) is poor. More detail about decision detection using DGMs will be presented in section 4.

Beyond decision detection, DGMs are used for labeling and segmenting sequences of observations in many different fields—including bioinformatics, ASR, Natural Language Processing (NLP), and information extraction. In particular, Dynamic Bayesian Networks (DBNs) are a popular model for probabilistic sequence modeling because they exploit structure in the problem to compactly represent distributions over multi-state and observation variables. Hidden Markov Models (HMMs), a special case of DBNs, are a classical method for important NLP applications such as unsupervised part-of-speech tagging (Gael et al., 2009) and grammar induction (Johnson et al., 2007) as well as for ASR. More complex DBNs have been used for applications such as DA recognition (Crook et al., 2009) and activity recognition (Bui et al., 2002).

Undirected graphical models (UGMs) are also valuable for building probabilistic models for segmenting and labeling sequence data. Conditional Random Fields (CRFs), a simple UGM case, can avoid the label bias problem (Lafferty et al., 2001) and outperform maximum entropy Markov models and HMMs.

However, the graphical models used in these applications are mainly non-hierarchical, including those in Bui et al. (2009). Only Sutton et al. (2007) proposed a three-level HGM (in the form of a dynamic CRF) for the joint noun phrase chunking and part of speech labeling problem; they showed that this model performs better than a non-hierarchical counterpart.

3 Data

For the experiments reported in this study, we used 17 meetings from the AMI Meeting Corpus[1], a freely available corpus of multi-party meetings with both audio and video recordings, and a wide range of annotated information including DAs and topic segmentation. The meetings last around 30 minutes each, and are scenario-driven, wherein four participants play different roles in a company’s design team: *project manager*, *marketing expert*, *interface designer* and *industrial designer*.

We use the same annotation scheme as Fernández et al. (2008) to model decision-making dialogue. As stated in section 2, this scheme distinguishes between a small number of DA types based on the role which they perform in the formulation of a decision. Besides improving the detection of decision discussions (Fernández et al., 2008), such a scheme also aids in summarization of them, because it indicates which utterances provide particular types of information.

The annotation scheme is based on the observation that a decision discussion typically contains the following main structural components: (a) A topic or issue requiring resolution is raised; (b) One or more possible resolutions are considered; (c) A particular resolution is agreed upon, and so adopted as the decision. Hence the scheme distinguishes between three main DDA classes: *issue* (*I*), *resolution* (*R*), and *agreement* (*A*). Class *R* is further subdivided into *resolution proposal* (*RP*) and *resolution restatement* (*RR*). *I* utterances introduce the topic of the decision discussion, examples being *“Are we going to have a backup?”* and *“But would a backup really be necessary?”* in Table 1. In comparison, *R* utterances specify the resolution which is ultimately adopted as the deci-

[1] http://corpus.amiproject.org/

(1) A: Are we going to have a backup? Or we do just–
B: But would a backup really be necessary?
A: I think maybe we could just go for the kinetic energy and be bold and innovative.
C: Yeah.
B: I think– yeah.
A: It could even be one of our selling points.
C: Yeah *–laugh–*.
D: Environmentally conscious or something.
A: Yeah.
B: Okay, fully kinetic energy.
D: Good.

Table 1: An excerpt from the AMI dialogue ES2015c. It has been modified slightly for presentation purposes.

sion. *RP* utterances propose this resolution (e.g. *"I think maybe we could just go for the kinetic energy . . . "*), while *RR* utterances close the discussion by confirming/summarizing the decision (e.g. *"Okay, fully kinetic energy"*). Finally, *A* utterances agree with the proposed resolution, signaling that it is adopted as the decision, (e.g. *"Yeah"*, *"Good"* and *"Okay"*). Unsurprisingly, an utterance may be assigned to more than one DDA class; and within a decision discussion, more than one utterance can be assigned to the same DDA class.

We use manual transcripts in the experiments described here. Inter-annotator agreement was satisfactory, with kappa values ranging from .63 to .73 for the four DDA classes. The manual transcripts contain a total of 15,680 utterances, and on average 40 DDAs per meeting. DDAs are sparse in the transcripts: for all DDAs, 6.7% of the totality of utterances; for *I*,1.6%; for *RP*, 2%; for *RR*, 0.5%; and for *A*, 2.6%. In all, 3753 utterances (i.e., 23.9%) are tagged as decision-related utterances, and on average there are 221 decision-related utterances per meeting.

4 Prior Work on Decision Detection using Graphical Models

To detect each individual DDA class, Bui et al. (2009) examined the four simple DGMs shown in Fig. 1. The DDA node is binary valued, with value 1 indicating the presence of a DDA and 0 its absence. The evidence node (E) is a multi-dimensional vector of observed values of non-lexical features. These include utterance features (UTT) such as length in words[2], duration in milliseconds, position within the meeting (as percentage of elapsed time), manually annotated dialogue act (DA) features[3] such as *inform*, *assess*, *suggest*, and prosodic features (PROS) such as energy and pitch. These features are the same as the non-lexical features used by Fernández et al. (2008). The hidden component node (C) in the *-mix* models represents the distribution of observable evidence E as a mixture of Gaussian distributions. The number of Gaussian components was hand-tuned during the training phase.

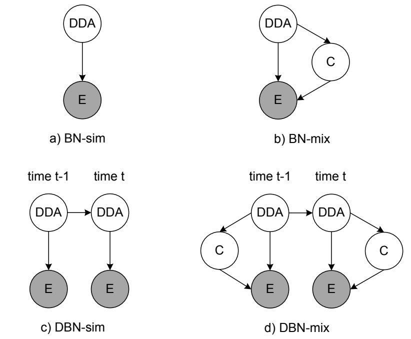

Figure 1: Simple DGMs for individual decision dialogue act detection. The clear nodes are hidden, and the shaded nodes are observable.

More complex models were constructed from the four simple models in Fig. 1 to allow for dependencies between different DDAs. For example, the model in Fig. 2 generalizes Fig. 1c with arcs connecting the DDA classes based on analysis of the annotated AMI data.

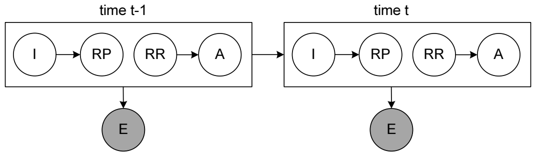

Figure 2: A DGM that takes the dependencies between decision dialogue acts into account.

Decision discussion regions were identified using the DGM output and the following two simple rules: (1) A decision discussion region begins with an *Issue* DDA; (2) A decision discussion region contains at least one *Issue* DDA and one *Resolution* DDA.

[2] This feature is a manual count of lexical tokens; but word count was extracted automatically from ASR output by Bui et al. (2009). We plan experiments to determine how much using ASR output degrades detection of decision regions.

[3] The authors used the AMI DA annotations.

The authors conducted experiments using the AMI corpus and found that when using non-lexical features, the DGMs outperform the hierarchical SVM classification method of (Fernández et al., 2008). The F1-score for the four DDA classes increased between 0.04 and 0.19 ($p < 0.005$), and for identifying decision discussion regions, by 0.05 ($p > 0.05$).

5 Hierarchical graphical models

Although the results just discussed showed graphical models are better than SVMs for detecting decision dialogue acts (Bui et al., 2009), two-level graphical models like those shown in Figs. 1 and 2 cannot exploit dependencies between high-level discourse items such as decision discussions and DDAs; and the "superclassifier" rule (Bui et al., 2009) used for detecting decision regions did not significantly improve the F1-score for decisions.

We thus investigate whether HGMs (structured as three or more levels) are superior for discovering the structure and learning the parameters of decision recognition. Our approach composes graphical models to increase hierarchy with an additional level above or below previous ones, or inserts a new level such as for discourse topics into the interior of a given model.

Fig. 3 shows a simple structure for three-level HGMs. The top level corresponds to high-level discourse regions such as decision discussions. The segmentation into these regions is represented in terms of a random variable (at each DR node) that takes on discrete values: {positive, negative} (the utterance belongs to a decision region or not) or {begin, middle, end, outside} (indicating the position of the utterance relative to a decision discussion region). The middle level corresponds to mid-level discourse items such as issues, resolution proposals, resolution restatements, and agreements. These classes ($C_1, C_2, ..., C_n$ nodes) are represented as a collection of random variables, each corresponding to an individual mid-level utterance class. For example, the middle level of the three-level HGM Fig. 3 could be the top-level of the two-level DGM in Fig. 2, each middle level node containing random variables for the DDA classes I, RP, RR, and A. The bottom level corresponds to vectors of observed features as before, e.g. lexical, utterance, and prosodic features.

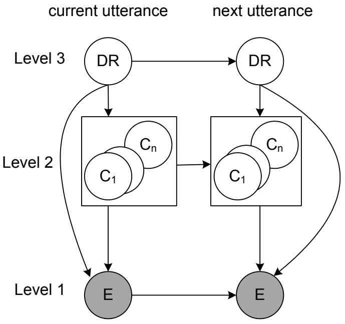

Figure 3: A simple structure of a three-level HGM: DRs are high-level discourse regions; $C_1, C_2, ..., C_n$ are mid-level utterance classes; and Es are vectors of observed features.

6 Experiments

The HGM classifier in Figure 3 was implemented in Matlab using the BNT software[4]. The classifier hypothesizes that an utterance belongs to a decision region if the marginal probability of the utterance's DR node is above a hand-tuned threshold. The threshold is selected using the ROC curve analysis[5] to obtain the highest F1-score. To evaluate the accuracy of hypothesized decision regions, we divided the dialogue into 30-second windows and evaluated on a per window basis.

The best model structure was selected by comparing the performance of various handcrafted structures. For example, the model in Fig. 4b outperforms the one in Fig. 4a. Fig. 4b explicitly models the dependency between the decision regions and the observed features.

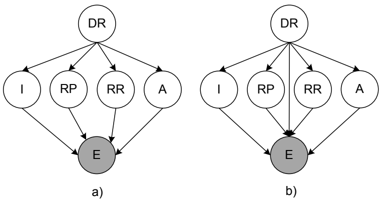

Figure 4: Three-level HGMs for recognition of decisions. This illustrates the choice of the structure for each time slice of the HGM sequence models.

Table 2 shows the results of 17-fold cross-validation for the hierarchical SVM classification (Fernández et al., 2008), rule-based classification with DGM output (Bui et al., 2009), and our HGM classification using the best combination of non-lexical features. All three methods

[4] http://www.cs.ubc.ca/∼murphyk/Software/BNT/bnt.html

[5] http://en.wikipedia.org/wiki/Receiver_operating_characteristic

were implemented by us using exactly the same data and 17-fold cross-validation. The features were selected based on the best combination of non-lexical features for each method. The HGM classifier outperforms both its SVM and DGM counterparts ($p < 0.0001$)[6]. In fact, even when the SVM uses lexical as well as non-lexical features, its F1-score is still lower than the HGM classifier.

Classifier	Pr	Re	F1
SVM	0.35	0.88	0.50
DGM	0.39	0.93	0.55
HGM	0.69	0.96	0.80

Table 2: Results for detection of decision discussion regions by the SVM super-classifier, rule-based DGM classifier, and HGM classifier, each using its best combination of non-lexical features: SVM (UTT+DA), DGM (UTT+DA+PROS), HGM (UTT+DA).

In contrast with the hierarchical SVM and rule-based DGM methods, the HGM method identifies decision-related utterances by exploiting not just DDAs but also direct dependencies between decision regions and UTT, DA, and PROS features. As mentioned in the second paragraph of this section, explicitly modeling the dependency between decision regions and observable features helps to improve detection of decision regions. Furthermore, a three-level HGM can straightforwardly model the composition of each high-level decision region as a sequence of mid-level DDA utterances. While the hierarchical SVM method can also take dependency between successive utterances into account, it has no principled way to associate this dependency with more extended decision regions. In addition, this dependency is only meaningful for lexical features (Fernández et al., 2008).

The HGM result presented in Table 2 was computed using the three-level DBN model (see Fig. 4b) using the combination of UTT and DA features. Without DA features, the F1-score degrades from 0.8 to 0.78. However, this difference is not statistically significant (i.e., $p > 0.5$).

7 Conclusions and Future Work

To detect decision discussions in multi-party dialogue, we investigated HGMs as an extension of the DGMs studied in (Bui et al., 2009). When using non-lexical features, HGMs outperform the non-hierarchical DGMs of (Bui et al., 2009) and also the hierarchical SVM classification method of Fernández et al. (2008). The F1-score for identifying decision discussion regions increased to 0.80 from 0.55 and 0.50 respectively ($p < 0.0001$).

In future work, we plan to (a) investigate cascaded learning methods (Sutton et al., 2007) to improve the detection of DDAs further by using detected decision regions and (b) extend HGMs beyond three levels in order to integrate useful semantic information such as topic structure.

Acknowledgments

The research reported in this paper was sponsored by the Department of the Navy, Office of Naval Research, under grants number N00014-09-1-0106 and N00014-09-1-0122. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research.

References

Satanjeev Banerjee, Carolyn Rosé, and Alex Rudnicky. 2005. The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In *Proceedings of the 10th International Conference on Human-Computer Interaction*.

H. H. Bui, S. Venkatesh, and G. West. 2002. Policy recognition in the abstract hidden markov model. *Journal of Artificial Intelligence Research*, 17:451–499.

Trung Huu Bui, Matthew Frampton, John Dowding, and Stanley Peters. 2009. Extracting decisions from multi-party dialogue using directed graphical models and semantic similarity. In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGdial09)*.

Nigel Crook, Ramon Granell, and Stephen Pulman. 2009. Unsupervised classification of dialogue acts using a dirichlet process mixture model. In *Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue*, pages 341–348.

Raquel Fernández, Matthew Frampton, Patrick Ehlen, Matthew Purver, and Stanley Peters. 2008. Modelling and detecting decisions in multi-party dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*.

[6] We used the paired t test for computing statistical significance. http://www.graphpad.com/quickcalcs/ttest1.cfm

Jurgen Van Gael, Andreas Vlachos, and Zoubin Ghahramani. 2009. The infinite HMM for unsupervised PoS tagging. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 678–687.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL).*

Daniel Gatica-Perez, Ian McCowan, Dong Zhang, and Samy Bengio. 2005. Detecting group interest level in meetings. In *Proceedings of ICASSP*.

Pey-Yun Hsueh and Johanna Moore. 2007. Automatic decision detection in meeting speech. In *Proceedings of MLMI 2007*, Lecture Notes in Computer Science. Springer-Verlag.

Adam Janin, Jeremy Ang, Sonali Bhagat, Rajdip Dhillon, Jane Edwards, Javier Marcías-Guarasa, Nelson Morgan, Barbara Peskin, Elizabeth Shriberg, Andreas Stolcke, Chuck Wooters, and Britta Wrede. 2004. The ICSI meeting project: Resources and research. In *Proceedings of the 2004 ICASSP NIST Meeting Recognition Workshop*.

Mark Johnson, Thomas Griffiths, and Sharon Goldwater. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 139–146, Rochester, New York, April. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann.

Iain McCowan, Jean Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. 2005. The AMI Meeting Corpus. In *Proceedings of Measuring Behavior, the 5th International Conference on Methods and Techniques in Behavioral Research*, Wageningen, Netherlands.

Matthew Purver, John Dowding, John Niekrasz, Patrick Ehlen, Sharareh Noorbaloochi, and Stanley Peters. 2007. Detecting and summarizing action items in multi-party dialogue. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium.

Charles Sutton, Andrew McCallum, and Khashayar Rohanimanesh. 2007. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research*, 8:693–723.

Steve Whittaker, Rachel Laban, and Simon Tucker. 2006. Analysing meeting records: An ethnographic study and technological implications. In S. Renals and S. Bengio, editors, *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Revised Selected Papers*, volume 3869 of *Lecture Notes in Computer Science*, pages 101–113. Springer.