The Predictive Power of Game Structure in Dialogue Act Recognition: Experimental Results Using Maximum Entropy Estimation

Massimo Poesio *Andrei Mikheev**

Human Communication Research Centre and Language Technology Group
University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, GB
http://www.hcrc.ed.ac.uk email: Massimo.Poesio@ed.ac.uk

ABSTRACT

Recognizing the dialogue act(s) performed by means of an utterance involves combining top-down expectations about the next likely 'move' in a dialogue with bottom-up information extracted from the speech signal. We compare two ways of generating expectations: one which makes the expectations depend only on the previous act, and one which also takes into account the fact that individual dialogue acts play a role as part of larger conversational structures ('games'). Our results indicate that exploiting game structure does lead to improved expectations.

1. INTRODUCTION

Recognizing the dialogue act(s) performed by means of an utterance involves combining top-down expectations about the next likely 'move' in a dialogue with bottom-up information extracted from the speech signal. The best current models of dialogue act recognition achieve an accuracy of about 70% on transcribed words and of 65% on recognized words (Stolcke et al., 1998; Reithinger and Klesen, 1997). We are trying to improve these results by finding better ways of exploiting top-down expectations about dialogue structure.

Nagata and Morimoto (1994); Reithinger and Klesen (1997); Poesio (1991) proposed to take advantage of expectations about what comes next in a conversation in order to improve dialogue act recognition. Statistical techniques to acquire this information were proposed by Nagata and Morimoto (1994) and Reithinger and Klesen (1997), among others. These researchers noted the resemblance of this task to that of part-of-speech tagging, and applied therefore techniques developed for this latter purpose, such as n-grams and hidden Markov models (Bahl et al., 1983; Katz, 1987), making the prediction of the next dialogue act depend on the previous n dialogue acts. Reithinger (1995) examined the predictive power of these models in isolation from bottom-up information, and found that they could achieve an accuracy of about 39%; he also noticed that bigrams performed as well as higher-order models (this result was confirmed in (Stolcke et al., 1998)). We likewise concentrated on the predictive power of expectations in isolation.

*Andrei Mikheev is now at Harlequin Ltd.

2. GAME-BASED EXPECTATIONS

It has been suggested in conversation analysis (Levinson, 1983) that both 'local' and 'global' organizing principles are at play in spoken conversations; these elements of organization include so-called ADJACENCY PAIRS (such as, e.g., QUESTION-ANSWER or INSTRUCT-ACCEPT) at the local level, as well as global organization principles such as, for example, the fact that phone conversations are normally opened by a signal from the person receiving the call, followed by the called introducing herself and the reason for the call, etc. It was also found that while these organization principles are not always respected, deviations from the norm tend to be marked–e.g., by pauses before rejecting a request. These findings led researchers to hypothesize that the participants in a conversation tend to act in accord with conventional routines called 'scripts' or GAMES (Carlson, 1983; Levin and Moore, 1978; Houghton and Isard, 1987) that generate expectations about what is coming next.

Although n-gram models can be viewed as a stripped-down version of game-based expectations, some information is lost when conversations are seen as simple sequences of moves - namely, the difference between *intra*-game and *inter*-game predictions. In a QUESTION-ANSWER game, for example, the initial QUESTION may be followed by a CLARIFY subgame. When this ends, the conversational participants return to the original question, and there is therefore an expectation that an ANSWER will follow–i.e., it is as if the subgame had not been there. (See Fig. 1.)

Also, whereas the (*intra*-game) prediction that a QUESTION will be followed by an ANSWER or a REQUEST FOR CLARIFICATION is fairly reliable, predicting what move will follow a move that closes a game is much harder, and if at all, it depends on even higher levels of structure. (E.g., the TRAINS conversations Gross et al. (1993); Heeman and Allen (1995) follow a rather predictable pattern whereby an initial phase in which the users ask various questions of the system, they then start making suggestions.)

One would expect a predictive technique that took these differences in account to fare better than one in which games are not treated as units. This hypothesis is indirectly supported by recent work on dialogue management, in which finite-state techniques for modeling intra-game behavior are combined with strategies depending on the task (e.g., which information is still needed) to infer inter-game behavior. Yet, we are not aware of previous work testing whether this was indeed the case. This has been the aim of our experiments.

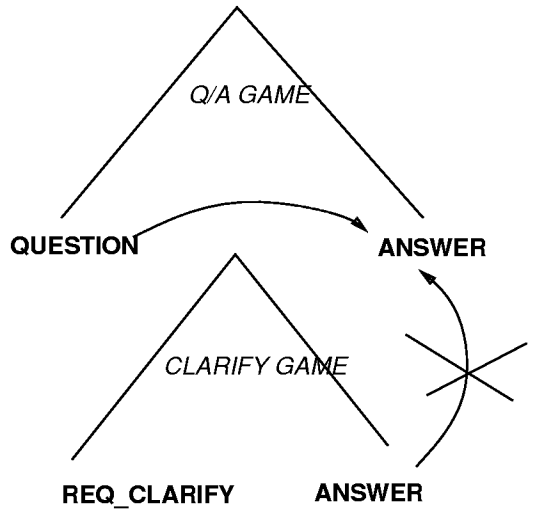

Figure 1: The expected effect of game structure: embedded games.

3. METHODS

We tested the hypothesis using the version of the MapTask corpus (Anderson et al., 1991) annotated for MOVES (corresponding to dialogue acts), games and TRANSACTIONS (Carletta et al., 1997). The MapTask corpus consists of 128 dialogues, for a total of 26,621 utterances. The classification scheme for moves includes 12 labels, 6 of which correspond to initiation acts that open a game (e.g., QUERY-W), whereas 6 correspond to response acts that respond to a previous move (e.g., REPLY-W). In addition to moves, their grouping into games, and the grouping of these into macro structures called transactions have also been annotated. We ran experiments comparing the predictive power of two types of statistical models trained on the MapTask corpus to perform dialogue act classification - models that take into account the notion of game, and models that don't.

Although several training techniques have been applied to the task of dialogue act recognition, no significant differences in performance have emerged so far. We used the maximum entropy estimation method of Berger et al. (1996) to build our language models, which offers a flexible way to integrate different types of knowledge sources and can be used to discover more complex features out of those encoded by hand. In HCRC, we developed tools that efficiently implement the technique (Mikheev, pear).

4. EXPERIMENTS

4.1. Baseline

The baseline with which to compare our results was determined by assigning to each utterance the most frequent dialogue act label, ACKNOWLEDGE. This results in a 20% accuracy. Randomly choosing an act results instead in 7% accuracy.

4.2. First Experiment - Bigrams

In our first experiment, we estimated the performance of bigram models in this corpus by assigning to each utterance a single input feature, the label of the previous act. (MOVE_LABEL) This made our model comparable to bigram models. We ran a 10-fold cross-training experiment. The results were: 38.6% first hypothesis correct (10,277 correct out of 26,621), 52% one of the first two hypotheses correct. These results are broadly comparable to those obtained with bigrams by Reithinger (1995), which confirms that maximum entropy estimation yields comparable results to n-grams when no other features are used. We also note that these results show that our corpus is interesting enough to get better results by prediction than by chance.

4.3. Second Experiment - Information about Games

In our second experiment, we still used the label of the previous move as an input feature, but in addition, we also assigned to each utterance two further input features: a first one, GAME_POSITION, specifying the position of the move in the game, with values INIT_GAME, IN_GAME and END_GAME, and a second feature GAME_TYPE specifying the type of game (QUESTION/ANSWER, INSTRUCT, etc.) for each move in the game except for the first (since in this case the move classification could have been derived directly from the game type). The experimental methodology used was as in the first experiment. The results: 50.63% first hypothesis correct, 67.07% one of the first two hypotheses was correct.

These results show that even taking into account the very simple notion of game structure encoded in the MapTask leads to an 12% improvement in accuracy (15% when the two best hypotheses are considered), *contra* Stolcke et al. (1998) (who, however, only took into account non-hierarchical structure information).

4.4. Third Experiment: Tracking Speaker Change

This accuracy can be improved even further by including a bottom-up input feature that is very easy to track, SPEAKER_CHANGE. Including this feature results in an accuracy of 54%.

We should notice that Chu-Carroll (1998) only gets a 4.4% improvement by adding speaker change to trigram information, which suggests that discourse structure information combines as well, or better, with other features than simple information about the previous dialogue act. In our own experiments, adding SPEAKER_CHANGE to MOVE_LABEL results in an accuracy of 41.8 % first hypothesis correct , 57.71% one of the first two correct.

4.5. Fourth Experiment - A more complex notion of game structure

In a fourth experiment, we compared the results obtained with the 'uniform' theory of games used in the MapTask with those obtained by separating task-oriented moves from dialogue control moves (ACKNOWLEDGE and CLARIFY), as suggested e.g., in (Poesio

and Traum, 1997). The results obtained in this case are 57.2% first hypothesis correct, 72.3% one of the first two hypotheses correct, once the SPEAKER_CHANGE feature is included.

5. RELATED WORK

(Stolcke et al., 1998) use a more complex set of labels (42 labels), but also get 35% accuracy by chance (50% for certain tasks) as opposed to 7% for us.

6. CONCLUSIONS

The experiments just discussed indicate that taking into account the hierarchical structure of a dialogue does result in significantly better predictions than simply keeping track of the previous move, even if only a simple notion of dialogue structure is considered. Two obvious questions to ask are whether the improvement in predictive power leads to improved performance overall, and whether we would still get improved predictions in case game structure were recognized out of the speech signal, instead of hand-coded.

The first question is easier to address, given that the MapTask has been completely transcribed, POS-tagged and parsed. As for the second question, we plan to start by trying to extract segmentation information (i.e., where games begin and end); the crucial problem to do this is how to extract prosodic information from the input, given that this information notoriously plays a crucial role in segmentation Nakatani et al. (1995).

ACKNOWLEDGMENTS

Massimo Poesio is an EPSRC Advanced Fellow. We wish to thank Jan Alexandersson, James Allen, Matthew Aylett, Chris Brew, Mark Core, Steve Isard, Daniel Jurafsky, David McKelvie, Steve Pulman, and Norbert Reithinger for comments and discussions, as well as audiences at the 1997 NGO workshop on Robust Processing of Dialogue in Nijmegen, at the HCRC's Dialogue Working Group, and at the University of Sheffield's Natural Language Group. Thanks also to the members of HCRC that collected and annotated the MapTask, without which this work would not have been possible.

REFERENCES

Anderson, A. H., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., and Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech*, 34(4):351–366.

Bahl, L. R., Jelinek, F., and Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190.

Berger, A., Della Pietra, S., and Della Pietra, V. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., and Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–32.

Carlson, L. (1983). *Dialogue Games*. D. Reidel, Dordrecht.

Chu-Carroll, J. (1998). A statistical model for discourse act recognition in dialogue interactions. In Chu-Carroll, J. and Green, N., editors, *Working Notes of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pages 12–17. AAAI Press.

Gross, D., Allen, J., and Traum, D. (1993). The TRAINS 91 dialogues. TRAINS Technical Note 92-1, Computer Science Dept. University of Rochester.

Heeman, P. A. and Allen, J. F. (1995). The TRAINS-93 dialogues. TRAINS Technical Note TN 94-2, University of Rochester, Dept. of Computer Science, Rochester, NY.

Houghton, G. and Isard, S. D. (1987). Why to speak, what to say, and how to say it. In Morris, P., editor, *Modeling Cognition*, pages 249–267. Wiley.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recogniser. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3):400–401.

Levin, J. A. and Moore, J. A. (1978). Dialogue games: Metacommunication strategies for natural language interaction. *Cognitive Science*, 1(4):395–420.

Levinson, S. (1983). *Pragmatics*. Cambridge University Press.

Mikheev, A. (To appear). Collocation lattices and maximum entropy models. To appear.

Nagata, M. and Morimoto, T. (1994). First steps towards statistical modeling of dialogue to predict the speech act type of the next utterance. *Speech Communication*, 15:193–203.

Nakatani, C., Hirschberg, J., and Grosz, B. J. (1995). Discourse structure in spoken language: Studies on speech corpora. In *Proc. AAAI Spring Symposium on Empirical Methods in Dialogue*, Stanford.

Poesio, M. (1991). Expectation-based recognition of discourse segmentation. In *Proc. AAAI Fall Symposium on Discourse Structure*, Asilomar, CA.

Poesio, M. and Traum, D. (1997). Conversational actions and discourse situations. *Computational Intelligence*, 13(3):309–347.

Reithinger, N. (1995). Some experiments in speech act prediction. In *Proc. of the AAAI Spring Symposium on Empirical Methods in Discourse*.

Reithinger, N. and Klesen, M. (1997). Dialogue act classification using language models. In *Proc. of Eurospeech-97*, pages 2235–2238, Rhodes.

Stolcke, A., Shriberg, E., Bates, R., Coccaro, N., Jurafsky, D., Martin, R., Meteer, M., Ries, K., Taylor, P., and Van-Ess-Dykema, C. (1998). Dialog act modeling for conversational speech. In *Working Notes of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pages 98–105. AAAI Press.