Dialogue Act Recognition with Bayesian Networks for Dutch Dialogues

Simon Keizer, Rieks op den Akker and Anton Nijholt
Department of Computer Science
University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands
{skeizer,infrieks,anijholt}@cs.utwente.nl

Abstract

This paper presents work on using Bayesian networks for the dialogue act recognition module of a dialogue system for Dutch dialogues. The Bayesian networks can be constructed from the data in an annotated dialogue corpus. For two series of experiments - using different corpora but the same annotation scheme - recognition results are presented and evaluated.

1 Introduction

In several papers (Nijholt, 2000; Luin et al., 2001; Nijholt et al., 2001) we reported on our virtual music centre - the VMC - a virtual environment inhabited by (embodied) agents and on multi-modal interaction between human users and these agents. Of these agents Karin is an embodied agent users can ask for information about theatre performances (see Figure 1).

A second agent is the navigation agent. Navigation is a) way finding - the user knows where he wants to go but doesn't know how to go there; or b) exploring the environment - the user walks through the environment to obtain an overview of the building and the objects, locations, rooms that are in it. Related to these navigation tasks the navigation assistant has the task to assist the visitor in a) explaining how to go from his current location to a location he is looking for and b) to give the agent information about objects, and locations in the environment. The navigation agent is not present as an avatar in the environment. The user sees the environment from a first person perspective and interacts with the agents by means of a Dutch dialogue. The user has two views of the environment: a) a first person view of the visible part of the 3D virtual theatre and b) an abstract 2D map of the floor of the building the user is visiting. This map is shown in a separate window. In a multi-modal interaction the user can point at locations or objects on the 2D map and either ask information about that object or location or he can ask the assistant to bring him to the location pointed at.

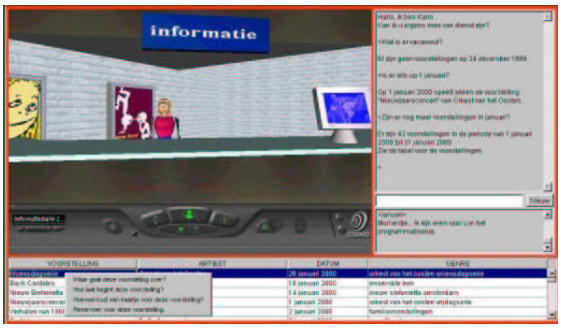

Figure 1: Karin in the VMC.

An important part of our dialogue systems for natural language interaction with agents is the module for recognition of the dialogue acts performed by the human user (visitor). This paper discusses the construction of and experiments with Bayesian networks as implementation of this module.

Various other work has been presented on using statistical techniques for dialogue act classification (Andernach, 1996; Stolcke et al., 2000), and even some first efforts on using Bayesian networks for this task (Pulman, 1996; Keizer, 2001). Other work on using Bayesian networks in dialogue systems aims more at interaction and user modelling (Paek and Horvitz, 2000) and does not specifically involve linguistic aspects.

The paper is organised as follows. Section 2

provides some necessary and general background about the use of Bayesian networks for speech act recognition. In Section 3 we discuss experiments with a Bayesian network for dialogue act classification based on a dialogue corpus for the Karin agent. In Section 4 we discuss our current experiments with a network for the navigation dialogue system that was automatically created from a small corpus. Section 5 reflects on our findings and presents plans for the near future.

2 Bayesian Networks and Speech Act Recognition

Since Austin and Searle deliberately producing a linguistic utterance ('locutionary act') is performing a speech act ('illocutionary act'). Many researchers have contributed in distinguishing and categorising types of speech acts we can perform. See (Traum, 2000) for a valuable discussion on dialogue act taxonomies and an extensive bibliography.

A dialogue system needs a user model. The better the user model the better the system is able to understand the user's intentions from the locutionary act. We consider the human participant in a dialogue as a source of communicative actions. Actions can be some verbal dialogue act or some non-verbal pointing act (the act of pointing at some object). We assume the user is rational: there is a dependency between the action performed and the intentional state of the user. If we restrict to communicative acts that are realized by uttering (speaking or typing) a sentence we can model the user by a probability distribution $P(U = u|DA = da)$: the probability that the user produces an utterance u (the stochastic variable U has value u) given that he performs a dialogue act da (DA has the value da). Or - maybe better: the confidence we can have in believing that the user uses utterance u if we know that the dialogue act he performs is da. Since there are many distinct wordings u for performing a given dialogue act da and on the other hand there are distinct dialogue acts that can be performed by the same utterance, we need more than superficial linguistic information to decide upon the intended dialogue act given an utterance. The task of the dialogue act recognition (DAR) module of a dialogue system is to answer the question: what is the most likely dialogue act da intended by the user given the system has observed the utterance u in a dialogue context c. (Notice that we have equated the utterance produced by the user with the utterance recognised by the system: there is no information loss between the module that records the utterance and the input of the dialogue act recognition module.)

To make this problem tractable we further restrict the model by assuming that a) the user engaged in a dialogue can only have the intention to perform one of a finite number of possible dialogue acts; b) each of the possible natural language utterances u produced by the user and observed by the system can be represented by a finite number of feature value pairs ($f_i = v_i$); and c) the dialogue context can be represented by a finite number of feature value pairs ($g_i = c_i$). Given these restrictions the DAR problem becomes to find that value da of DA that maximises

$$P(\ DA = da \mid f_1 = v_1, \ldots, f_n = v_n, \\ g_1 = c_1, \ldots, g_m = c_m\ ).$$

For the probabilistic model from which this can be computed we use a Bayesian network (Pearl, 1988). A Bayesian network is a directed acyclic graph in which the nodes represent the stochastic variables considered, while the structure (given by the arcs between the nodes) constitutes a set of conditional independencies among these variables: a variable is conditionally independent of its non-descendants in the network, given its parents in the network. Consider the network in Figure 2: it contains one node representing the dialogue act (DA), 3 nodes representing utterance features ($NumWrds$, $CanYou$ and $IWant$) and a node representing a context feature ($PrevDA$). From the network structure follows that for example variable DA is conditionally independent of variable $NumWrds$, given variable $CanYou$.

The conditional independencies make the model computationally more feasible: finding a specification of the joint probability distribution (jpd) for the model reduces to finding the conditional probability distributions of each of the variables given their network parents. In our example network, the following jpd specification holds:

$$P(DA, NumWrds, CanYou, IWant, PrevDA) = \\ P(IWant) \cdot P(PrevDA|DA) \cdot P(CanYou) \cdot$$

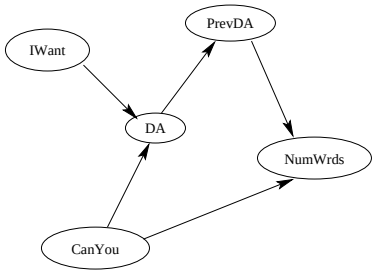

Figure 2: A Bayesian Network for Dialogue Act Recognition.

$$\cdot P(DA|CanYou, IWant) \cdot$$
$$\cdot P(NumWrds|CanYou, PrevDA)$$

The construction of a Bayesian network hence amounts to choosing a network structure (the conditional independencies) and choosing the conditional probability distributions. In practice, the probabilities will have to be assessed from empirical data by using statistical techniques. The structure can generated from data too, but another option is to choose it manually: the arcs in the network can be chosen, based on the intuition that they represent a causal or temporal relationship between two variables. Strictly spoken however, a Bayesian network only represents informational relationships between variables.

Notice that the machine learning technique known as Naive Bayes Classifier (see for instance (Mitchell, 1997)) assumes that all variables are conditionally independent of each other given the variable that has to be classified. A Naive Bayes classifier can be seen as a special case of a Bayesian network classifier, where the network structure consists of arcs from the class variable to all variables representing the features: see Figure 3.

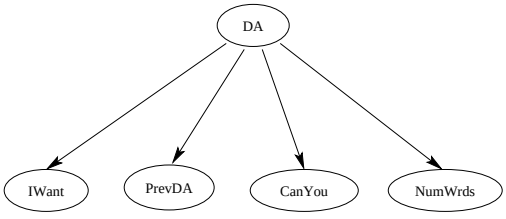

Figure 3: Naive Bayes classifier as Bayesian Network.

Naive Bayes classifiers will perform as good as the Bayesian network technique only if indeed all feature variables are conditionally independent, given the class variable. The problem is of course how do we know that they are conditionally independent? If we don't have complete analytical knowledge about the (in)dependencies, only analysing the data can give an answer to this question. The advantage of using Bayesian networks is that methods exist to construct the network structure as well as the conditional probabilities. Moreover Bayesian networks are more flexible in their use: unlike Bayesian classifiers we can retrieve the posterior probabilities of all the network variables without re-computation of the model. The same advantage do Bayesian networks have over Decision Tree learning methods like C4.5 that output a decision tree for classifying instances with respect to a given selected class variable. Experiments have shown that Naive Bayesian classifiers give results that are as good as or even better than those obtained by decision tree classification techniques. Hence, there are theoretical as well as practical reasons to use Bayesian networks. However, since there is hardly any experience in using Bayesian networks for dialogue act classification we have to do experiments to see whether this technique also performs better than the alternatives mentioned above for this particular application.

The next two sections describe experiments with 1) the SCHISMA corpus - elaborating on previous work described in (Keizer, 2001) - and 2) a preliminary small corpus of navigation dialogues. We motivate our choice of dialogue acts and features and present some first results in training a Bayesian network and testing its performance.

3 Experiments with the Schisma corpus

3.1 Dialogue acts and features

The current dialogue system for interacting with Karin is based on analyses of the SCHISMA corpus. This is a corpus of 64 dialogues, obtained through Wizard of Oz experiments. The interaction between the wizard - a human simulating the system to be developed - and the human user was established through keyboard-entered utterances, so the dialogues are textual. The task at hand is information exchange and transaction: users are enabled to make inquiries about theatre per-

formances scheduled and if desired, make ticket reservations.

We have manually annotated 20 dialogues from the SCHISMA corpus, using two layers of the DAMSL multi-layer annotation scheme (Allen and Core, 1997), a standard for annotating task-oriented dialogues in general. The layer of Forward-looking Functions contains acts that characterise the effect an utterance has on the subsequent dialogue, while acts on the layer of Backward-looking Functions indicate how an utterance relates to the previous dialogue. Because DAMSL does not provide a refined set of dialogue acts concerning information-exchange, we have added some new dialogue acts. For example, `ref-question`, `if-question` and `alts-question` were added as acts that further specify the existing `info-request`.

For the experiments, we selected a subset of forward- and backward-looking functions from the hierarchy that we judged as the most important ones to recognise: those are listed in Table 1. In Figure 4, a fragment of an example dialogue between S (the server) and C (the client) is given, in which we have indicated what forward- and backward-looking functions were performed in each utterance.

Forward-looking Functions	**Backward-looking Functions**
assert	accept
open-option	approve
request	reject
ref-question	disapprove
if-question	hold
alts-question	acknowledge
action-directive	not-understood
offer	positive answer
commit	negative answer
conventional	feedback
expressive	otherbf
otherff	

Table 1: Dialogue Acts for SCHISMA.

The user utterances have also been tagged manually with linguistic features. We have distinguished the features in Table 2, assuming they can be provided for by a linguistic parser.

The dialogue context features selected include the backward-looking function of the last system utterance and the forward-looking function of the previous user utterance. In the experiment with

S: Hello, how can I help you?
`conventional`
C: When can I see Herman Finkers?
`ref-question`; `otherbf`
S: On Saturday the 12th at 20h.
`assert`; `pos-answer`
C: I would like 2 tickets please.
`action-directive`; `otherbf`
S: Do you have a discount card?
`if-question`; `hold`
...

Figure 4: Dialogue fragment with forward- and backward-looking functions.

Sentence Type
declarative
yn-question
wh-question
imperative
noun phrase
adverbial
adjective
number
interjective
continuation

Subject Type
first person
second person
third person

Punctuation
period
question mark
exclam. mark
comma
none

Table 2: Utterance features for SCHISMA.

the SCHISMA dialogues we have constructed a network structure (see Figure 5) by hand and then used the data of the annotated dialogues to train the required conditional probabilities.

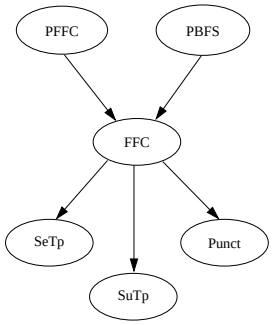

Figure 5: Bayesian network for DAR to be trained with the SCHISMA dialogues.

The choice of structure is based on the intuition that the model reflects how a client decides which communicative action to take; although the arcs themselves have no explicit meaning - they only contribute to the set of conditional independencies - they can be seen here as a kind of temporal or causal relationships between the variables (as mentioned earlier in Section 2): *given* the dialogue context -

defined by the previous forward-looking function of the client ($PFFC$) and the previous backward-looking function of the server ($PBFS$), the client decides which forward-looking function to perform (FFC); from this decision he/she formulates a natural language utterance with certain features including the sentence type ($SeTp$) the subject type ($SuTp$) and punctuation ($Punct$).

Recalling the notion of conditional independence in Bayesian networks described in Section 2, it follows that by choosing the network structure of Figure 5, we have made the (admittedly, disputable) assumption that, given the forward-looking function of the client, the three utterance features are conditionally independent of each other.

3.2 Results and evaluation

For assessing the conditional probability distributions, we have used the Maximum A Posteriori (MAP) learning technique - see e.g. (Heckerman, 1999). For training we have used 330 data samples which is 75% of the available data; the remaining samples have been used for testing. We have measured the performance of the network in terms of the accuracy of estimating the correct forward-looking function for different cases of available evidence, varying from having no evidence at all to having evidence on all features. This resulted in an average accuracy of 43.5%. Adding complete evidence to the network for every test sample resulted in 38.7% accuracy.

As the amount of data from the SCHISMA corpus currently available is rather small, the results cannot expected to be very good and more data have to be collected for further experiments. Still, the testing results show that the accuracy is significantly better than an expected accuracy of 8.3% in the case of guessing the dialogue act randomly. A tighter baseline commonly used is the relative frequency of the most frequent dialogue act. For the data used here, this gives a baseline of 32.5%, which is still less than our network's accuracy.

4 Experiments with the navigation corpus

4.1 Dialogue acts and features

A small corpus of dialogues was derived from the first implementation of a dialogue system for interaction with the navigation agent. For the experiments with the navigation corpus we also use the DAMSL layers of Forward- and Backward-looking functions. On each of these two layers we only distinguish dialogue acts on the first level of the hierarchies (see Table 3 for the dialogue acts used); a more refined subcategorisation should be performed by a second step in the DAR module. The dialogue acts in Table 1 can be found at the deeper levels of the DAMSL hierarchy, e.g. a `request` is a special case of a `infl_addr_fut_act` and an `acknowledge` is a special case of an `understanding`. The dialogue act recogniser may also use more application specific knowledge in further identification of the user intention. Information that may be used is dialogue information concerning topic/focus.

Forward-looking Functions
`statement`
`infl_addr_fut_act`
`info_request`
`comm_sp_fut_act`
`conventional`
`expl_performative`
`exclamation`

Backward-looking Functions
`agreement`
`understanding`
`answer`

Table 3: Dialogue Acts for Navigation.

For the navigation dialogues, we have chosen a set of surface features of what will eventually be spoken utterances, in contrast to the typed dialogues in the SCHISMA corpus. Therefore, we don't use a textual feature like punctuation. For each utterance, the feature values are found automatically using a tagger (the features in the SCHISMA dialogues were tagged manually). In Table 4 we have listed the features with their possible values we initially consider relevant.

The dialogue context features include the backward- and forward-looking function of the previous dialogue act. This is always a dialogue act performed by the system. The possible dialogue acts performed by the system are the same as those performed by the user.

The network is generated from data that were obtained by manually annotating the user utterances in the navigation corpus following the DAMSL instructions as close as possible. As with every categorisation there are problematic

Features	Values
lenq	one, few, many
iswh	true, false
not_in_prev	true, false
startsWithCanYou	true, false
startsWithCanI	true, false
startsWithIWant	true, false
containsPositive	true, false
containsNegative	true, false
containsOkay	true, false
containsLocativePrep	true, false
containsLocativeAdverb	true, false
containsTell	true, false
containsDo	true, false

Table 4: Surface features of user utterances and their possible values.

border cases, e.g. when to annotate with indirect speech acts. We used the criterion that such an act should be recognised without task-specific considerations. Therefore the utterance "I want to make a phone-call" is annotated as a `statement` although eventually it should be interpreted as an `info_request` ("where can I find a phone?") in the context of a navigation dialogue.

After the dialogue act has been recognised the navigation agent will make a plan for further actions and perform the planned actions. We will not discuss that here.

4.2 Results and evaluation

In this experiment the data are used for learning both structure and conditional probabilities of a Bayesian network. We have used an implementation of the K2 algorithm (Cooper and Herskovits, 1992) to generate the network structure and then - like in the SCHISMA experiment - used MAP to assess the conditional probability distributions.

Starting from the small corpus of navigation dialogues, a procedure has been planned to iteratively enlarge the corpus: given the annotated corpus, derive a network, use the network in a dialogue system, test the network and add these dialogues - with the corrected backward- and forward-looking functions - to the corpus. This results in a more extended set of annotated dialogues. And we start again. After each of the cycles we compare the results (in terms of accuracies) with the results of the previous cycle. This should give more insight in the usefulness of the features and values chosen for the Bayesian network. After deciding to adapt the set of features we automatically annotate the corpus; we derive a new network and we test again.

The current corpus is too small to expect good results from a generated network, especially if the data are used for learning both the structure and the probability distributions. From the initial corpus of 81 utterances 75% was used for generating a Bayesian network. Testing on the remaining 25% resulted in accuracy of 57.1% for classifying the forward-looking function and 81.0% for classifying the backward-looking function. After this first cycle, new data have been generated interactively, following the procedure described above. The Bayesian network trained from this new data set resulted in the improved accuracies of 76.5% and 88.2% for classifying the forward- and backward-looking function respectively. Following this training and testing procedure, we hope to develop Bayesian networks with increasing performance.

5 Discussion and conclusions

In this paper we have discussed the use of Bayesian networks for dialogue act recognition in a natural language dialogue system. We have described the construction of such networks from data in two cases: 1) using annotated dialogues from the SCHISMA corpus - information exchange and transaction - and 2) using a small corpus of annotated navigation dialogues.

As the amount of data currently available is rather small (especially the navigation corpus), the network performances measured are not too impressive. In order to get more data, we have developed a testing environment which at the same time enables us to enlarge the corpus. With the increasing amount of data we hope to construct Bayesian networks with increasing performance. As for the SCHISMA corpus, there are 44 dialogues that remain to be annotated, also resulting in more data.

One of the first and most important questions to be answered concerns the selection of a set of features (and their values) that set up the model. We started with a set of features selected on intuition. Then the dialogue corpus was annotated. As a result of experiments we may conclude that some of the features have no selective value, so

we can leave them out of the model.

In the future we would like to compare the approach of using Bayesian networks with other classifiers that can also be constructed from data, e.g. decision trees or Bayesian classifiers. Figure 5 shows the accuracies of three different classifiers that were generated from the current set of navigation data.

Class variable	Bayesian network	Decision tree	Naive Bayes
forw_funct	76.5%	50.0%	55.9%
backw_funct	88.2%	64.7%	61.8%

Table 5: Accuracies of three different classifiers for classifying the forward-looking function (forw_funct) and backward-looking function (backw_funct), where all classifiers have been built from the same set of navigation data.

In our future experiments we will take into account more refined performance measures like precision and recall and confusion matrices in which classification results for individual dialogue act types are shown. Such results can help us make decisions w.r.t. the selected dialogue act types and features.

Furthermore, non-verbal communicative actions like pointing at objects in the virtual environment could be relevant in recognising dialog acts and should therefore be made available as possible features in our Bayesian network classifiers.

Acknowledgement

We would like to thank the referees for their comments on our paper; these have been very useful to us in preparing this final version.

References

J. Allen and M. Core. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers. URL: `http://www.cs.rochester.edu/research/trains/annotation`.

T. Andernach. 1996. A machine learning approach to the classification and prediction of dialogue utterances. In *Proceedings of the Second International Conference on New Methods in Language Processing (NeMLaP-2)*, pages 98–109, Ankara, Turkey.

G. F. Cooper and E. Herskovits. 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

D. Heckerman. 1999. A tutorial on learning with Bayesian networks. In M. Jordan, editor, *Learning in Graphical Models*. MIT Press, Cambridge MA.

S. Keizer. 2001. A Bayesian approach to dialogue act classification. In P. Kühnlein, H. Rieser, and H. Zeevat, editors, *BI-DIALOG 2001: Proc. of the 5th Workshop on Formal Semantics and Pragmatics of Dialogue*, pages 210–218.

J. van Luin, R. op den Akker, and A. Nijholt. 2001. A dialogue agent for navigation support in virtual reality. In J. Jacko and A. Sears, editors, *ACM SIGCHI Conf. CHI 2001: Anyone. Anywhere*, pages 117–118, Seattle. Association for Computing Machinery.

T.M. Mitchell. 1997. *Machine Learning*. Computer Science Series. McGraw-Hill.

A. Nijholt, J. Zwiers, and B. van Dijk. 2001. Maps, agents and dialogue for exploring a virtual world. In N. Callaos, S. Long, and M. Loutfi, editors, *5th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2001)*, volume VII of *Human Information and Education Systems*, pages 94–99, Orlando, July.

A. Nijholt. 2000. Multimodal interactions with agents in virtual worlds. In N. Kasabov, editor, *Future Directions for Intelligent Systems and Information Science*, Physica-Verlag: Studies in Fuzziness and Soft Computing, chapter 8, pages 148–173. Springer.

T. Paek and E. Horvitz. 2000. Conversation as action under uncertainty. In *16th Conference on Uncertainty in Artificial Intelligence (UAI-2000)*, pages 455–464, San Francisco, CA, June. Morgan Kaufmann Publishers.

J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

S.G. Pulman. 1996. Conversational games, belief revision and Bayesian networks. In J. Landsbergen, J. Odijk, K. van Deemter, and G. Veldhuijzen van Zanten, editors, *Computational Linguistics in the Netherlands*.

A. Stolcke et al. 2000. Dialogue act modelling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

D.R. Traum. 2000. 20 questions on dialogue act taxonomies. *Journal of Semantics*, 17(1):7–30.