Natural Language Engineering **13** (4): 287–316. © 2006 Cambridge University Press
doi:10.1017/S1351324905004067 First published online 15 February 2006 Printed in the United Kingdom

Dialogue act recognition under uncertainty using Bayesian networks

SIMON KEIZER

Computational Linguistics & Artificial Intelligence Group, Faculty of Arts, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: s.keizer@uvt.nl

RIEKS OP DEN AKKER

Human Media Interaction Group, Department of Electrical Engineering, Mathematics and Computer Science, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
e-mail: infrieks@cs.utwente.nl

(*Received 8 September 2004; revised 18 April 2005*)

Abstract

In this paper we discuss the task of dialogue act recognition as a part of interpreting user utterances in context. To deal with the uncertainty that is inherent in natural language processing in general and dialogue act recognition in particular we use machine learning techniques to train classifiers from corpus data. These classifiers make use of both lexical features of the (Dutch) keyboard-typed utterances in the corpus used, and context features in the form of dialogue acts of previous utterances. In particular, we consider probabilistic models in the form of Bayesian networks to be proposed as a more general framework for dealing with uncertainty in the dialogue modelling process.

1 Introduction

The research described in this paper is concerned with the problem of uncertainty due to incomplete information for interpreting natural language utterances in the context of a dialogue. In particular, we considered the task of recognising dialogue acts, a concept for characterising the communicative action as a contribution within a coherent dialogue. A well-known technique for dealing with uncertainty but one that has not been explored to a great extent in NLP yet is the use of Bayesian networks, a special kind of probabilistic model. Exceptions are for example (Pulman, 1996; Black *et al*., 2003), work on multimodal interaction (Heckerman and Horvitz, 1998; Bilvi and Pelachaud, 2003), and recently also work on multiparty conversations in meetings (Galley *et al*., 2004). We think that Bayesian networks form a promising technique for use in dialogue systems, and our experiments on dialogue act classification confirm that hypothesis, although much further research lies ahead. In Keizer (2003) the use of Bayesian networks for dialogue act classification is discussed as part of a more general idea of using Bayesian networks to deal with uncertainty in dialogue systems.

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

In the particular experiments that were set up, two research questions were asked: 1) what is the performance of the Bayesian network classifiers in comparison to other classifier types examined, and 2) which information is most relevant for determining the dialogue act type of a given utterance, where we restricted ourselves to information in terms of lexical features of the textual utterances in the corpus and context features in terms of dialogue acts performed in previous utterances.

In Section 2 we discuss the notion of dialogue acts as originated from Searle's speech act theory. Formally, a dialogue act is defined as a pair consisting of a *dialogue act type* and a *semantic content*; in Section 3 we describe the dialogue act types that we have subjected to our dialogue act classification experiments. In Section 4, we discuss the uncertainty underlying this task and in Section 5 we describe a probabilistic approach to this uncertainty in the form of Bayesian Networks and explain how they can be used as dialogue act classifiers. Then the actual experiments are described (Section 6) and the results are presented (Section 7). Things are wrapped up with some conclusions and future recommendations in Section 8.

2 Dialogue acts

Imagine a situation in which the visitor of a virtual building addresses the virtual receptionist, asking:

(1) "is there a phone that I can use?"

Now, the receptionist, i.e., the (navigation) agent with underlying dialogue system, has to analyse this utterance and come up with an interpretation based on the information that was extracted from the utterance by other, lower-level components (syntactic features, occurrence of certain words, word patterns, aspects of intonation, etc.) and not without importance, information about the context (which also includes non-linguistic aspects, such as gestures). This interpretation is to be used in deciding which next action to take, in particular, which utterance to produce as a meaningful response to the visitor:

(2) "yes, down the hall, on your left."

However, we first need to point out that utterances should be seen primarily as *communicative acts* in a cooperative process between two dialogue participants. Concerning the more specific case of natural language utterances we use the term *speech acts*, a notion that originates from theories in language philosophy, especially from John Searle's speech act theory (Searle, 1969).

Austin (1975) had already pointed out that producing a natural language utterance has several dimensions, or rather, perspectives, to it: he particularly distinguished the locutionary, the illocutionary and the perlocutionary act that may be performed in the event of an utterance.

The *locutionary act* involves the physical act of producing certain sounds, thereby producing words from a certain vocabulary and complying with a certain grammar, and thus producing a phrase or sentence with a certain sense and reference. However,

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

IN bringing about this locutionary act, the speaker may also be performing an *illocutionary act*, i.e., an act that expresses the communicative force of the utterance. In producing the utterance in 1, the speaker asks a `question`, or, performs an `information request`. One may paraphrase the utterance as "I hereby ask you if there is a phone that I can use", hence making explicit the communicative force. The above paraphrase is an example of a *performative sentence*, the kind of utterances that Austin was interested in originally, although he aimed at more typical performatives such as "I hereby name this ship the Queen Elizabeth" or "you are fired". However, also an utterance like "a man is walking through the park" has an illocutionary force, namely that of `asserting` or `claiming`: "I hereby assert that a man is walking through the park".

The *perlocutionary act* is the act that is performed BY producing the utterance. In such an act the effect of the utterance is expressed, for example, by saying "a man is walking through the park" the speaker may be threatening the addressee besides asserting something to him. As opposed to the illocutionary act of asserting however one cannot paraphrase the utterance into "I hereby threaten you that a man is walking through the park", and therefore, the threatening is not an illocutionary, but a perlocutionary act.

Searle elaborated further on the illocutionary acts as introduced by Austin and concluded that speakers produce such speech acts by means of so-called *illocutionary force indicating devices* (IFIDs) that are provided for in a natural language. The illocutionary force (IF) indicates the type of the illocutionary act, for example, the IF of the utterance of "go to hell" could be `directive`, and has a *semantic content* (Searle used the term 'propositional content') given by the suggested future action of the addressee, namely, the going to hell. By using the elements of a language in a certain way, i.e., by complying to certain silent, though conventional rules, the speaker can bring about a certain speech act. He may for example use the directive mood to produce a directive act, not by saying "you are going to hell" or "are you going to hell?", but by saying "go to hell".

Searle's general hypothesis was that "speaking a language is engaging in a rule-governed form of behaviour". He tried to find these rules underlying the performance of illocutionary acts by means of IFIDs and the very formal nature of this investigation has been influential to the idea from the technical perspective to develop models in which the relationship between illocutionary acts and IFIDs – more generally, the relation between speech acts and linguistic material – is specified and to use those models in language technology, e.g., in dialogue systems.

Most dialogue systems that support relatively complex dialogues (involving complex tasks, mixed initiative, multimodality, etc.) use some concept of dialogue act modelling: user utterances are interpreted in terms of a dialogue act type DAT from a predefined set of types and a semantic content SC that can be derived from syntactic and semantic analysis of the utterance. The dialogue act (DAT, SC) found is used to update the internally represented dialogue state and to decide on a reaction.

In this paper we will concentrate on the interpretation part, and in particular, the task of determining what dialogue act type to assign to the utterance, using

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

the information made available to us (from other components in the system, e.g., a part-of-speech tagger).

3 Dialogue act types

As already indicated in the previous section, a complete dialogue act consists of a dialogue act type DAT and a semantic content SC. The question is now what dialogue act types one can distinguish and how they are organised. Austin and Searle already did early attempts at a categorisation of possible speech act types. Many refinements, improvements and alternatives to this categorisation have followed since. Examples of recent attempts at categorisations can be found elsewhere (Bunt, 1995; Bunt, 1996; Poesio and Traum, 1997; Poesio and Traum, 1998; Alexandersson *et al*., 1998; Carletta *et al*., 1996); Traum (2000) gives a good overview.

The dialogue act types we used for our experiments were taken from a dialogue act hierarchy that was based on the hierarchy underlying the DAMSL annotation scheme (Allen and Core, 1997), which was intended as a standard for marking up task-oriented dialogues. This dialogue act hierarchy consists of several layers each of which contains a set of hierarchically organised dialogue act types. Each layer covers a different aspect of the communicative action that is performed in an utterance. Our dialogue act classification experiments involve only two layers in our scheme, namely, the Forward-looking Functions (FLFs) and the Backward-looking Functions (BLFs). FLFs characterise the way in which an utterance influences the future dialogue, while BLFs characterise the way in which an utterance relates to parts of the dialogue history. Hence, each utterance may be labelled with both a FLF and a BLF (except for special cases like opening utterances that have no BLF). In the following outline of the dialogue act types on these two levels in our scheme, 'S' refers to the speaker and 'H' refers to the hearer. The examples are taken from the theatre domain, just like the dialogues in the corpus we used for the experiments (see Section 6.1).

Forward-looking Functions

- `statement`: S makes a claim about the world.
  - `assert`: S tries to change the belief of H.
    "'Tommie' begint om 20:00u" *(Eng.: Tommie starts at 20:00h)*; also elliptical sentences such as "ja" *(Eng.: yes)*, that can be paraphrased as "ja, ik heb een kortingskaart" *(Eng.: yes, I have a discount card)*.
  - `reassert`: S thinks the claim has already been made.
  - `other`
- `influencing-addressee-future-action`: S is suggesting potential actions to H, beyond answering a request for information.
  - `action-directive`: S is creating an obligation that H do the action unless H explicitly indicates otherwise.
    "ik wil twee kaartjes" *(Eng.: two tickets please)*; "doe maar 2 mei" *(Eng.: make it the 2nd of May)*.
  - `open-option`: S is not creating this obligation.
    "vanavond kunt u naar Herman Finkers in de Grote Zaal" *(Eng.: this evening you can see Herman Finkers performing in the Grote Zaal)*.

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

- `info-request`: S is requesting (by just asking or in another, indirect way) H for information.
  1. **query-if**: the speaker asks the hearer whether something is the case or not. "heeft u een reductiekaart?" *(Eng.: do you have a discount card?)*.
  2. **query-ref**: the speaker asks the hearer for information in the form of references that satisfy some specification given by the speaker. "welke voorstellingen zijn er vanavond?" *(Eng.: which performances are on this evening?)*.
- `committing-speaker-future-action`: S is potentially committing to intend to perform a future action.
  - `offer`: S's commitment is contingent on H's agreement.
    "zal ik de kaartjes voor u reserveren?" *(Eng.: shall I book the tickets for you?)*.
  - `commit`: otherwise.
    "ik zal de kaartjes voor u reserveren" *(Eng.: I will book the tickets for you)*.
- `conventional`:
  - `opening`: S summons H and/or starts the interaction.
  - `closing`: S closes the dialogue or is dismissing H.
    "bedankt" *(Eng.: thanks)*; "tot ziens"
- `explicit-performative`
- `exclamation`
- `other-forward-function`

Backward-looking Functions

- `agreement`: S is addressing a previous proposal, request, or claim.
  - `accept`: S is agreeing to all of the proposal, request, or claim.

    context: "zal ik de kaartjes voor u reserveren?" *(Eng.: shall I book the tickets for you?)*

    "ja graag" *(Eng.: yes please)*
  - `accept-part`: S is agreeing only partly to the proposal, request, or claim.
  - `reject`: S is disagreeing to all of the proposal, request, or claim.
    (using the same context as for `accept`)

    "nee dank je" *(Eng.: no thanks)*
  - `reject-part`: S is disagreeing only partly to the proposal, request, or claim.
  - `hold`: S is not stating his attitude towards the proposal, request, or claim; the speaker postpones his response to a request, proposal or claim of the hearer. In the example below, the speaker does not directly respond to the hearer's request with an accept or reject, but he starts a sub-dialogue to obtain more information.

    context: "ik wilde graag reserveren voor 5 personen" *(Eng.: I would like to make reservations for 5 people)*

    De Grote Zaal heeft de volgende rangen: . . . *(Eng.: The Grote Zaal has the following classes/circles: . . .)*
  - `maybe`: S is stating his attitude, but is non-committal to the proposal, request, or claim.
- `understanding`: utterances concerning the understanding between S and H, ranging from merely hearing the words to fully identifying intentions.
  - `signal-non-understanding`: S asks H what H said or meant (not to be confused with cases of hold in which also a request for clarification is made, but rather a request for additional information; it does not signal non-understanding).

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

— `signal-understanding`: S explicitly signals understanding (it may occur simultaneously with a signal-non-understanding however).
  – `acknowledge`: S signals understanding, but not necessarily acceptance.
  – `repeat-rephrase`: in order to signal understanding.
  – `completion`: S signals understanding by finishing (part of) the utterance H is making.

— `correct-misspeaking`: S adds a correction, indicating that he believes that H has not said what he actually intended.

- **answer**: the speaker responds to an `info_request` of the hearer.
  — **positive-answer**:

    context: "welke voorstellingen zijn er vanavond?" *(Eng.: what performances are on this evening?)*

    "Vanavond kunt u naar: ..." *(Eng.: For this evening, the following performances have been scheduled: ...)*

    – **confirm**: positive response to a `query_if`.

  — **no-answer-feedback**: the speaker indicates he cannot give the information requested.
  — **negative-answer**:

    context: "welke voorstellingen zijn er vanavond?" *(Eng.: what performances are on this evening?)*

    "Vanavond zijn er helaas geen voorstellingen" *(Eng.: Unfortunately, no performances have been scheduled for this evening)*

    – **disconfirm**: negative response to a `query_if`.

- `information-relation`: this is a tag intended to capture how the (semantic) content of this utterance relates to the content of its antecedent (for example, various tags for rhetorical relations could be classified under this tag).

4 Dialogue act recognition

The task of dialogue act recognition (DAR) involves finding an interpretation of a given utterance u in the context of a dialogue in terms of a dialogue act with type dat and semantic content sc:

$$u \longrightarrow \boxed{\text{DAR}} \longrightarrow (dat, sc)$$

In our experiments we concentrate on determining which of the dialogue act types from the set of types used, say dat_k, should be associated with a given utterance-in-context u represented by a set of features $f_1, \ldots, f_m$ and their respective values $v_1, \ldots, v_m$. The types in our case are selections from the hierarchies of Forward- and Backward-looking Functions as described in Sections 3 and 6.2. We refer to this task as dialogue act classification (DAC):

$$(f_1 = v_1, \ldots, f_m = v_m) \longrightarrow \boxed{\text{DAC}} \longrightarrow dat_k$$

The features may concern both features of the utterance itself, to be derived from the output of components for speech analysis and syntactic/semantic analysis, and features of the context such as the dialogue act types of previous utterances.

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

4.1 Uncertainty

The model that describes the domain at hand, i.e., natural language dialogue, is in general incomplete: because of the infinite spectrum of relevant phenomena in natural language, some phenomena will inevitably have to be ignored (or: will be overlooked) in the model. This is why the relationships between the various elements of the model (variables, propositions, concepts) can never be determined with absolute certainty.

Furthermore, in a particular situation one may have to deal with incomplete information in determining the state of affairs represented through the model. For example, in a spoken dialogue system, the result of speech recognition is always incomplete (i.e., the speech recogniser delivers partial information) and causes the further interpretation of an utterance (in terms of a dialogue state update) to be inherently uncertain. In the particular case of dialogue act classification some of the features in the model may not be available – or their values are themselves uncertain – and this makes the classification result uncertain. For example, a POS tagger may yield several alternative readings for an utterance, which has consequences for any utterance features derived from that result with regard to their certainty.

In using a finite set of dialogue act types, many subtleties in the use of an utterance may be abstracted away, thereby causing uncertainties in the interpretation of the linguistic behaviour of a human user who is not aware of this abstraction. Wittgenstein (1958) already observed that there are infinitely many ways of using language, and in that sense any set of dialogue act types that characterise different kinds of language usage is incomplete. However, if the search for classification is restricted to an inventory of performative verbs as Austin did, or to illocutionary acts as Searle did, the incompleteness may still be manageable.

The problem of uncertainty can be approached along two dimensions (besides the use of verification strategies within the dialogue for increasing the confidence in the interpretation of utterances).

1. *Representation*: One can deal with uncertainty by making it explicit through the assignment of qualifications of uncertainty to the elements – propositions – in the model. In particular, we can use an uncertainty measure like probability – numbers between 0 and 1 – and use probability calculus to combine the probabilities of different propositions.
2. *Construction*: Because even expert knowledge is insufficient to be certain about the models constructed from that knowledge, we can use learning techniques to construct models from raw data automatically, or to refine/improve existing models with new data.

Bayesian networks are precisely the kind of model that uses probability theory and that can be constructed by using machine learning from both raw data and expert knowledge.

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

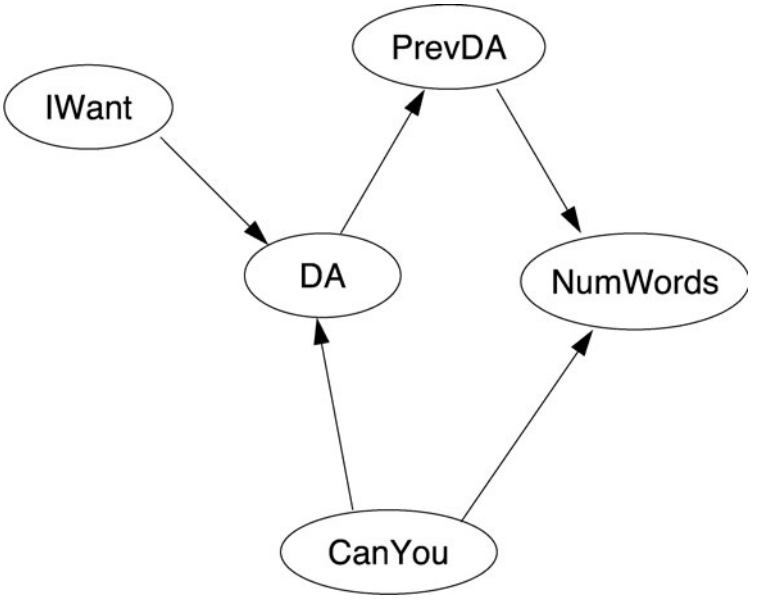


Fig. 1. Example of a Bayesian network.

4.2 Machine learning

By using raw data we can construct models for a domain in which general relationships are represented and use them for prediction or for analysing particular states of the domain. From a dialogue corpus we can construct models in which for example surface features of utterances and dialogue act types are specified and use these models to analyse and interpret new (user) utterances in a new particular dialogue.

In the application of machine learning, two choices need to be made (given that relevant data are available and ready to use): 1) what kind of model, which sort of representation, will be used, and 2) which technique, which learning algorithm, will be used to construct models from the data.

In the experiments that have been done, 3 classifier types have been examined: besides of course Bayesian networks, also Decision Trees and Naive Bayes models have been trained and evaluated.

5 Bayesian networks

Bayesian networks (Pearl, 1988) are probabilistic models in which a set of conditional independencies have been made explicit by means of a directed acyclic graph. The independencies make both constructing such models and doing inference with them more easy. One needs less data for learning and inference algorithms can employ the independencies. Figure 1 shows an example of a Bayesian network.

More formally, a Bayesian network consists of a set of propositional variables $X_1, \ldots, X_n$ that characterise the phenomena in terms of which we want to reason about the domain; the probabilistic model is then given by a directed acyclic graph

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

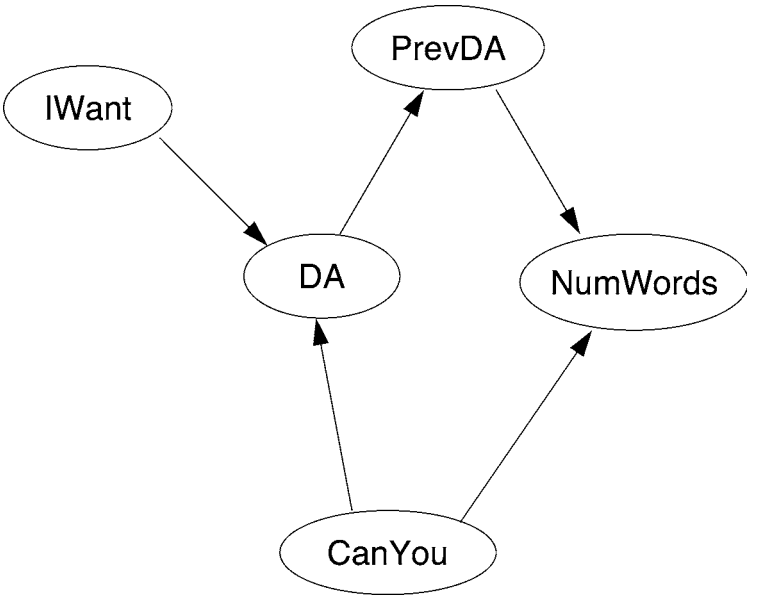


Fig. 2. A Bayesian network for dialogue act recognition.

and a set of conditional probability distributions. The nodes $V_1, \ldots, V_n$ in the graph each correspond uniquely with one variable in the model; the arcs between the nodes represent dependencies between the variables, or rather, the graph structure stipulates a set of conditional independencies according to the graph theoretical principle of 'd-separation'. For example, one of the independence relationships represented in the example Bayesian network in Figure 1, is that *DA* is conditionally independent of *NumWrds*, given that the values of *CanYou* and *PrevDA* are known.

For a complete specification of the probabilistic model given these independencies, only a conditional probability distribution is required for each of the variables given the variables corresponding to the parent nodes of that variable in the graph. If for example, V_k and V_l are the parents of V_m, then a conditional probability distribution $P(X_m|X_k, X_l)$ is required. The graph hence provides for a factorisation of the joint probability distribution (jpd):

$$P(X_1, \ldots, X_n) = \prod_i P(X_i|pa(X_i)) \tag{1}$$

where $pa(X_i)$ denotes the variables corresponding to the parents of node V_i in the graph (if a node V_k has no parents, the corresponding factor in the jpd is an unconditional $P(X_k)$). For the network of Figure 2 we get the following jpd:

$$\begin{aligned} P(DA, NumWrds, CanYou, IWant, PrevDA) = \\ P(IWant) \cdot P(CanYou) \cdot P(DA|CanYou, IWant) \cdot \\ P(PrevDA|DA) \cdot P(NumWrds|CanYou, PrevDA) \end{aligned}$$

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

5.1 Classification

Suppose we obtain new information about our model, i.e. we obtain evidence $E = e$, the instantiation of a set of variables E in the network. Then our posterior belief in $X = x$ (the instantiation of a different set of variables) can be obtained from the joint probability distribution as follows:

$$P(X = x|E = e) = \frac{P(X = x, E = e)}{P(E = e)} = \frac{\sum_{s} P(X = x, E = e, S = s)}{\sum_{x,s} P(X = x, E = e, S = s)} \quad (2)$$

where S refers to all variables excluding X and E. This process is called *belief updating*. In the process of what is called *belief revision*, we may determine the configuration y^* of the variables in a subset Y, that maximises the joint probability over these variables, given the evidence e:

$$y^* = \underset{y}{argmax}\ P(Y = y|E = e) \quad (3)$$

The Bayesian network in Figure 1, which has been reproduced in Figure 2, may be used for dialogue act classification. The network contains one node representing the dialogue act type of a given utterance (this is the class variable *DA*), three nodes that represent surface features of that utterance (*NumWords*, *CanYou* and *IWant*) and one node that represents a feature of the context in which the utterance was produced, the dialogue act type of the previous utterance (*PrevDA*). Classification using this Bayesian network operates in a straightforward manner, using the maximisation specified in (3), with the simplification that evidence is available for all feature variables, i.e., all variables except the class variable, and the most probable class value is found by maximising over the values of the single class variable.

However, in contrast to many other classification techniques it is not required that all feature values are given. If an NLP module in the dialogue system indicates a low confidence with respect to the value of a feature, it may be decided to leave the corresponding variable in the network uninstantiated. Similarly, it is not required that a Bayesian network contains only one class variable. Several aspects in the interpretation of a user utterance that need to be computed after NLP, such as dialogue acts in different dimensions or references that have to be resolved, may be modelled by several (non-evidence) variables in a single model, thereby allowing the modelling of any dependencies between them. After entering available information, these aspects can be queried for in terms of the most probable combination of the uninstantiated variables.

In a Naive Bayes classifier, the features are considered conditionally independent, given the class variable. Such a model however can also be expressed in a Bayesian network: see Figure 3. Therefore, we may conclude that in theory, Bayesian network classifiers can be found that perform at least as good as Naive Bayes classifiers.

5.2 Learning

Bayesian networks can both be constructed by means of expert knowledge and from raw empirical data through machine learning techniques. Given a set of propositional

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

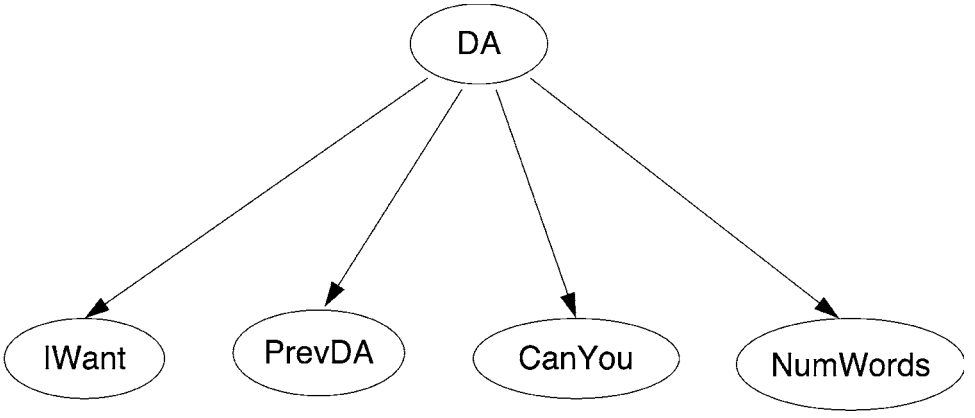


Fig. 3. Naive Bayes classifier as Bayesian network.

variables for describing the domain, both finding a network structure and assessing the numbers that specify the conditional probability distributions can be done by using experts and also from raw data; combinations are quite common, for example, using expert statements about conditional independencies, a graph structure can be found and subsequently, the appropriate probability tables can be assessed from data. In learning the probability tables, one can also start from initial numbers obtained from expert knowledge.

The Bayesian network classifiers used in the experiments have been constructed by using the K2-algorithm (Cooper and Herskovits, 1992). Given an ordering on the variables $X_1, \ldots, X_n$, the graph structure is built incrementally by selecting for each node V_i a set of parents from its predecessors $V_1, \ldots, V_{i-1}$ in the ordering. This selection is based on a statistical measure on the quality of a graph, relative to the data (we used an entropy measure as implemented in the machine learning tool that we made use of, see Section 7). The numbers for the required conditional probability distributions are obtained in a straightforward manner from relative frequencies.

6 Experiments

The experiments are concerned with the classification of dialogue acts by applying machine learning techniques to corpus data. There have been many research efforts in which models are induced from the data in an annotated corpus (Nagata and Morimoto, 1994; Samuel *et al.*, 1998; Kipp, 1998; Wright, 1998; Wright *et al.*, 1999; Stolcke *et al.*, 2000); also, there is ongoing work on dialogue act classification in the analysis of multiparty conversations (Shriberg *et al.*, 2004). The use of Bayesian networks however is fairly new. In Pulman (1996), some preliminary experiments on using Bayesian networks for dialogue move recognition are presented, and in Black *et al.* (2003) the classification experiments also involved Naive Bayes classifiers (a special form of Bayesian networks, see Section 5). Our first explorations on using Bayesian networks for dialogue act recognition and some preliminary experiments have been presented in Keizer (2001a, 2001b, 2002).

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

The experiments presented here have been subdivided into classifying the backward-looking function (class variable **blf**) and forward-looking function (class variable **flf**) of client (user) utterances. The features we use concern both utterance features obtained from the output of a part-of-speech (POS) tagger, and context features, involving the dialogue act types of previous utterances. In some other empirical approaches to dialogue act recognition in which use has been made of very large corpora, selecting the most plausible dialogue act type was merely based on sequences of words and additional information such as prosodic cues or dialogue history (Reithinger and Klesen, 1997; Stolcke *et al.*, 2000). Because our corpus is relatively small, a large number of features will make it impossible to achieve results that are statistically significant enough and therefore, the use of higher-level information to arrive at a relatively small set of linguistic features is more appropriate. Besides that, our goal of finding meaningful linguistic features that are informative for dialogue act recognition would have to be put aside in the case of using merely word sequences. Furthermore, we are now able to do an exhaustive examination of all possible subsets of the features and in this way find features that yield the best results in terms of classifier performance.

6.1 The data

We have made use of data from a corpus of 65 Dutch, keyboard-typed, mixed-initiative dialogues that was obtained through Wizard of Oz experiments. The dialogues concern information-exchange about theatre performances scheduled for an existing Dutch theatre; users could ask for information about the performances and possibly make reservations for tickets. The corpus dialogues were collected for use in the development of a dialogue agent to operate in a virtual environment called the Virtual Music Center (VMC), next to other agents such as a navigation agent (Nijholt, 2000; van Luin *et al.*, 2001; Nijholt *et al.*, 2001; Hofs *et al.*, 2003). The corpus has been annotated with dialogue act types from a hierarchy based on the DAMSL annotation scheme, see also Section 3. To get an idea of the dialogues, a small fragment of a dialogue from the corpus is given, together with the FLF and BLF (in that order) assigned in the annotation. The `no-blf` tag indicates that no backward-looking aspect can be identified in the utterance, for example, in utterances at the beginning of a dialogue.

C : *Wanneer draait Sweeney Todd?* (Eng.: When is Sweeny Todd on?)
[`query-ref,no-blf`]

S : *U kunt "Sweeney Todd" zien in de Grote Zaal in de periode 28 december 1993 t/m 30 december 1993.* (Eng: You can see Sweeny Todd in the 'Grote Zaal' in the period December 28 to 30 1993.)
[`open-option,positive-answer`]

C : *En Under a blue roof?* (Eng.: What about Under a blue roof?)
[`query-ref,no-blf`]

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

Table 1. *Frequencies of the backward_looking_function tags*

Tag	Client	Server	Total
backward_looking_function	920	1127	2047
`accept`	69	54	123
`reject`	13	7	20
`hold`	39	161	200
`signal_non_understanding`	3	20	23
`signal_understanding`	7	0	7
`positive_answer`	162	399	561
`negative_answer`	30	42	72
`no_answer_feedback`	3	63	66
`no_blf`	589	379	968
`other`	5	2	7

S : *U kunt "Under a blue Roof" zien in de Grote Zaal op 19 mei.* (Eng.: You can see 'Under a blue Roof' in the 'Grote Zaal' on May 19.)
[`open-option,positive-answer`]

C : *Kan ik daar kaarten voor reserveren* (Eng.: Can I order tickets for that)
[`query-if,no-blf`]

S : *Heeft u een reductiekaart?* (Eng.: Do you have a reduction card?)
[`query-if,hold`]

S : *Hoeveel kaartjes wilt U?* (Eng.: How many tickets do you require?)
[`query-ref,no-blf`]

C : *Ik heb geen reductiekaart.* (Eng.: I don't have a reduction card.)
[`assert,negative-answer`]

C : *Vier kaartjes graag.* (Eng.: Four tickets please.)
[`action-directive,positive-answer`]

6.2 The class values

From the hierarchies of backward- and forward-looking functions, we selected two sets of dialogue act types as class-values for classification. In Tables 1 and 2, these selection are given, together with their frequency in the corpus, both for client and server utterances. Each of the 65 dialogues in the corpus are divided up into turns, utterances and segments, the last of which are the units for dialogue act annotation. From the uppermost rows in both of the tables one can see that the corpus consists of 2047 segments, each tagged with a forward- and a backward-looking function; 920 of the segments are associated with the client (the user) and the remaining 1127 are associated with the server (the wizard simulating the dialogue system). The selections were based on some initial intuition concerning what types a dialogue system should primarily distinguish between for dialogues in the domain of information-exchange and transaction. At the same time, the number of classes was kept limited for

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

Table 2. *Frequencies of the forward_looking_function tags*

Tag	Client	Server	Total
forward_looking_function	920	1127	2047
`statement`	123	694	817
`action_directive`	239	11	250
`open_option`	0	111	111
`query_if`	70	38	108
`query_ref`	433	165	598
`committing_speaker_future_action`	0	70	70
`conventional`	29	31	60
`other_forward_function`	19	7	26
`other`	7	0	7

technical reasons. Therefore, types with very low frequency in the corpus for both client and server utterances were not selected.

The selection procedure ensured that no two types could be chosen that are engaged in a supertype-subtype relation, ruling out the combination of, for example, `accept` and `agreement`. Furthermore, in each selection an additional class value with label 'other' was introduced, covering the remaining dialogue act types from the original hierarchy that were not yet covered by the initial selection.

6.3 Utterance features

The surface features of utterances are computed dynamically during the process of generating suitable data-files from the corpus for specific machine learning experiments. For computing these features the output of a simple POS tagger is used. Based on an initial intuition on what features could be informative with respect to identifying the forward- and backward-looking function (flf and blf respectively) of an utterance, restricted by the information from the available tagger, we defined an initial set of utterance features.

- **lenq**: qualitative account of the number of words in the utterance; 'one' for 1 word, 'few' for 2, 3 or 4 words, and 'many' for 5 or more words. Can be an indicator for short answers like "yes" or "no" consisting of only one word; for short utterances consisting of 4 words at most, like "nee bedankt", "4 kaartjes graag" or "van wie is Othello?".
- **isContinue**: 'true' if the utterance shows a 'continuation pattern', i.e. starts with the (Dutch) word "en" ("and…", "what about…"), or "maar" ("but…"). This feature might be a good indicator if we also take into account the previous flf of the same speaker: if an utterance shows a continuation pattern, it will more likely have the same flf as the previous utterance.
- **startsWithWHExpr**: 'true' if the utterance starts with a 'wh-expression', i.e. either a wh-word ("wie", "wat", "welke", etc.) or a preposition followed by a wh-word ("voor wie", "op wat"). Clearly, this feature is a strong indicator for the syntactic sentence type 'wh-question' and this is expected to be closely related to the flf-type `info-request` and `query-ref` in particular.
- **endsWithQuestionMark**: 'true' if the utterance ends with a question mark. This feature is particularly expected to separate the flf-types `statement` and `info-request`.

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

Utterances of type `action-directive` can either occur as 'questioning' ("wilt u voor me reserveren?") or not ("ik wil reserveren"). This feature might also be relevant in distinguishing utterances of type `offer` from those of type `commit`, although the present experiments do not include this distinction (both types merge into the more general type `committing-speaker-future-action`).

- **startsWithCanYou**: 'true' if the utterance starts with phrases like "kunt u", "kun je", "zou je", "wilt u". This feature seems a good indicator for identifying requests: `action-directive`, `open-option`, `query-if` and `query-ref`.
- **startsWithCanI**: 'true' if the utterance starts with any of the phrases "kan ik", "mag ik", "moet ik". This feature seems a good indicator for a `query-if` or `action-directive`.
- **startsWithIWant**: 'true' if the utterance starts with any of the phrases "ik wil", "ik zou", "ik moet". This feature is also meant for identifying requests.
- **containsPositive**: 'true' if the utterance contains a word like "ja", "jazeker", "inderdaad". ("yes", "yes, indeed", "exactly", etc.) This feature should be informative with respect to identifying positive responses: `positive-answer` and `accept`.
- **containsNegative**: 'true' if the utterance contains words like "nee", "niet", "geen". ("no", "not", "none", etc.) This feature should be informative with respect to identifying negative responses: `negative-answer` and `reject`.
- **containsOkay**: 'true' if the utterance contains a word like "ok", "akkoord", "ja", "goed". ("ok", "alright", etc.) This feature should also be informative with respect to identifying positive responses, especially the blf's of type `accept`.
- **containsTell**: 'true' if the utterance contains a verb like "vertel", "zeg", "noemen". ("tell", "say", "name") Utterances containing verbs like these – particularly in questions – are more probably information requests than statements for example.
- **containsDo**: 'true' if the utterance contains a word like "breng", "doe", "brengen". ("take", "bring", etc.) These verbs clearly point into the direction of the utterance being an `action-directive`.
- **containsLocativePrep**: 'true' if the utterance contains a locative preposition like "van", "naar", "langs". ("from", "to", "via", etc.)
- **containsLocativeAdverb**: 'true' if the utterance contains a locative adverb like "hier", "daar", "voor". ("here", "there", etc.)

The last two features are actually 'relics' from earlier experiments with a very small collection of made-up dialogues between a user and a navigation agent. These features were chosen with the specific underlying task of navigation in mind and seem to be of less relevance concerning the dialogues in the corpus we use. In the subsequent models we have replaced these features with their 'temporal' equivalents (see Section 7.1).

6.4 Context features

In our classification model, the context is represented by features of previous parts of the dialogue. In the annotations the dialogues are divided up into *turns*, alternating between client and server turns. Each turn however consists of one or more *utterances* and each utterance consists of one or more *segments*, which are the basic units to assign FLFs and BLFs and utterance features to. So in order to take into account context we generated sequences of segments – each last one of them associated with a client turn – in preparing the datasets for training. The number of segments in a sequence is measured by the *window-size*, i.e., a window-size of 1 means that we only consider the current segment and do not take into account any context features

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

and a window-size of 2 means that we also take into account the features of the preceding segment.

The choice for at least taking into account dialogue act types in the context is motivated by the notion of *adjacency pairs* (Schegloff and Sacks, 1973; Goffman, 1981), suggesting that for example a `question` is a strong indicator for an `answer` to follow. The two utterances 1 and 2 in Section 2 form an example of an adjacency pair. To keep the number of features limited, we did not take any utterance features of previous segments into account.

For the context features we used generic feature names: the forward- and backward-looking functions of the *i*th segment preceding the current segment are referred to as **flf-i** and **blf-i** respectively. So, if a window-size of 2 is taken, the context features are **flf-1** and **blf-1**. For all context features included, one additional value 'null' had to be introduced to account for instances in which a previous segment did not exist – for example, the first segment in a dialogue does not have any previous segments at all.

In our initial model, sequences of segments are generated from n-grams of turns. In this way, it could be ensured that the current segment stemmed from a client turn and preceding segments alternately stemmed from server and client turns respectively (in Section 7.2, a different approach to generating sequences of segments is presented). To deal with the fact that a turn may consist of more than one segment, we chose to generate all possible sequences of segments from a single n-gram of turns, thereby accepting the additional noise as a consequence.

6.5 Discussion

Before we turn to the presentation of the results of the experiments, some remarks need to be made regarding the appropriateness of the experimental setting. The aim of the classification experiments is to use the (Bayesian network) classifiers within a dialogue system. In the current experimental setting however, several assumptions and design choices have been made that have consequences for application in a dialogue system. The corpus consists of text-based dialogues, making the results less appropriate for use in a spoken dialogue system. Spoken dialogues have different linguistic characteristics, and the additional and very relevant information regarding intonation and prosody has to be taken into account. Another issue is the segmentation of the turns in a dialogue into segments. As the segments are defined as the basic units of dialogue act assignment with no explicit syntactic ground, and multiple segments within a single turn are not rare, the segmentation is still a non-trivial process that has to be accounted for in a dialogue system when using classifiers that are trained according to the current experimental setting.

Furthermore, we have to be aware of the fact that in each dialogue situation, a classifier will assume the previous dialogue acts to be known (as feature-value pairs). The fact however is that the previous dialogue acts are obtained from previous classifications and are not as certain as feature-value pairs suggest. (In a way, this is also the case with utterance features, which may also be taken from the partial information from an NLP module.) One possible solution of avoiding this kind of accumulative error is the use of a so-called *Dynamic Bayesian Network*, in which

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

information on the previous dialogue acts would be incorporated in terms of (prior) probability distributions – obtained from a previous classification – instead of the instantiation of variables. In this setting it is also possible that the (posterior) probability distribution of the variables representing previous dialogue acts may be different, based on new information (about the now current utterance) after belief updating. It also provides a method, well-founded in probability theory, in which the fact that a classification at one moment may turn out to be wrong later on and needs to be revised, is accounted for. For the moment however, we had to leave any research on using Dynamic Bayesian networks in dialogue modelling for future work.

7 Results

The experiments presented here are first explorations in finding appropriate models for dialogue act classification, which includes both finding relevant utterance and context features, and comparing different types of classifiers and machine learning techniques, with a particular focus on Bayesian networks. The experiments consist of three series, every other series containing some modifications to the model used in the preceding series. In Section 6 we presented the initial model; in the second model the way in which context is represented has been modified (Section 7.2), and finally, some further modifications have been made to both model and the set-up of the experiments (Section 7.3).

From the annotated corpus, six different datasets were created for each model. First, we created a dataset for each of the three window-sizes 1, 2 and 3 (see Section 6.4). After that, from each of those three datasets separate datasets for classification of the backward-looking function (with class variable **blf**) and the forward-looking function (with class variable **flf**) were produced: in the dataset for **blf**-classification the **flf** had been removed, and in the dataset for **flf**-classification the **blf** had been removed.

For each of these datasets, 10-fold cross-validation of three different classifiers was performed for all possible subsets of the features specified. From the evaluation results the feature selection with the highest classification accuracy was derived for each of the three classifiers: the Naive Bayes classifier, the Bayesian Network classifier, and the Decision Tree classifier. In the setup of our testing environment, we made use of the Java-classes of the WEKA toolbox for machine learning implementations (Witten and Frank, 1999).

7.1 Initial model

In Tables 3 and 4 we have summarised the accuracy results which are given in percentages with in addition 95% confidence intervals. The results show that in all cases, all classifiers show a significant improvement w.r.t. the chosen *baseline*, i.e., the relative frequency of the most frequent class value. Because of the high frequency of `no-blf` in the corpus, the baselines for blf-classification become much higher in comparison to the average relative frequency of $1/10 = 10\%$. The accuracies for the different classifiers do not differ significantly, although the Naive Bayes classifier performs worst in 5 out of 6 cases. This may be explained by the general property of Naive Bayes classifiers performing worse when applied to sparse data.

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

Table 3. *Maximum accuracies for blf-classification in the Initial Model*

Context	Naive Bayes	Bayes Net	Decision Tree	Baseline
1-grams	70.89 (±3.06)	71.01 (±3.06)	70.65 (±3.07)	64.38
2-grams	82.09 (±2.25)	83.44 (±2.19)	82.81 (±2.22)	54.37
3-grams	81.83 (±2.17)	84.15 (±2.06)	83.24 (±2.10)	54.91

Table 4. *Maximum accuracies for flf-classification in the Initial Model*

Context	Naive Bayes	Bayes Net	Decision Tree	Baseline
1-grams	62.01 (±3.27)	66.98 (±3.17)	67.10 (±3.17)	46.75
2-grams	69.13 (±2.72)	70.93 (±2.67)	71.83 (±2.65)	40.59
3-grams	68.95 (±2.61)	71.43 (±2.54)	70.85 (±2.56)	40.05

The accuracies increase significantly when including blf and flf of the previous segment: when going from 1-grams to 2-grams, blf-classification gains more than 10% for all 3 classifiers. When including blf and flf from the previous *client* turn (i.e. in the case of 3-grams), the classification accuracies do not change significantly ((Black *et al*., 2003) have presented similar results concerning the window-sizes). The accuracy of the Naive Bayes classifier has even decreased for both flf- and blf-classification; the Decision Tree classifier only gains accuracy for blf-classification; the Bayes Net is the only classifier that shows improved accuracy for both flf- and blf-classification.

The resulting feature selections did not give a very clear and unambiguous view of which features are most significant for classification. Nevertheless, there are some useful observations when comparing the various selections to be mentioned. In Table 5 a summarisation of the selection results is presented. Each row gives for a particular feature the number of times it occurred in an optimal feature selection, in the different cases – determined by the window size (1, 2 or 3) and the class variable (flf or blf). As we are considering 3 three different classifiers, these occurrences are maximally three times per case. For example, the feature **endsWithQuestionMark** occurred in two of the three selections in the case of blf-classification and in all three selections in the case of flf-classification, making 5 occurrences in total. To put it another way, two of the three classifiers came up with **endsWithQuestionMark** in their feature selection that resulted in the highest accuracy for blf-classification with a window-size of 1, and all three classifiers came up with that feature in the case of flf-classification with a window-size of 1.

Let us first consider the results based on only utterance features, i.e., the window-size is 1. Here, the features **containsNegative** and **containsOkay** are in all optimal selections, while **startsWithCanYou** and **containsLocativeAdverb** are in none of the selections.

In general, the pattern of occurrences of the features is maintained when context features are added. The features **containsNegative** and **containsOkay** are very

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

Table 5. *Number of times the features are in the optimal selections for the 3 classifier types*

	Window-size 1			Window-size 2			Window-size 3			
	blf	flf	**total**	blf	flf	**total**	blf	flf	**total**	**total**
blf-2							1	1	**2**	
flf-2							1	3	**4**	
blf-1				3	3	**6**	3	1	**4**	
flf-1				3	3	**6**	3	3	**6**	
lenq	2	3	**5**	2	3	**5**	3	3	**6**	16
isContinue	0	3	**3**	2	2	**4**	1	3	**4**	11
startsWithWHExpr	0	3	**3**	3	3	**6**	3	3	**6**	15
endsWithQuestMark	2	3	**5**	3	3	**6**	2	3	**5**	16
startsWithCanYou	0	0	**0**	1	1	**2**	0	0	**0**	2
startsWithCanI	0	2	**2**	1	3	**4**	0	3	**3**	9
startsWithIWant	2	1	**3**	3	3	**6**	2	3	**5**	14
containsPositive	1	3	**4**	2	3	**5**	2	3	**5**	14
containsNegative	3	3	**6**	3	3	**6**	3	2	**5**	17
containsOkay	3	3	**6**	3	3	**6**	3	3	**6**	18
containsLocatPrep	1	2	**3**	3	0	**3**	0	0	**0**	6
containsLocatAdv	0	0	**0**	1	1	**2**	0	1	**1**	3
containsTell	1	0	**1**	0	1	**1**	2	2	**4**	6
containsDo	2	3	**5**	0	3	**3**	2	2	**4**	12

frequent, where one might add the additional frequently occurring features **lenq**, **startsWithWHExpr** and **endsWithQuestionMark**. The features **startsWithCanYou** and **containsLocativeAdverb** are in only a few of the selections.

Now let us consider the differences in the selection results between those for blf-classification and those for flf-classification. In general, we can say that for flf-classification, more utterance features are included in optimal selections than for blf-classification. For example, if we consider the window-size 1 case, the features **isContinue** and **startsWithWHExpr** are in none of the selections for blf-classification, but in all of the selections for flf-classification when considering window-size 1. Concerning **startsWithWHExpr**, this confirms our hypothesis that it would be a significant feature for distinguishing the flf `info-request` from other flf-types. One reasonable explanation of the fact that **isContinue** is specifically significant for flf-classification, even without taking into account the previous flf of the same speaker as suggested in Section 6.3, might be that continuations in utterances are often used for (continuing their) `info-requests` and much less frequently for other flf-types.

When taking into account the flf and blf of the previous (system) utterance (a window-size of 2), besides the features that were in all selections for the case of window-size 1, also the features **startsWithWHExpr**, **endsWithQuestionMark** and **startsWithIWant** are in all classifier selections. Where **endsWithQuestionMark** was already in all but one selection in the window-size 1 case, the significance of **startsWithWHExpr** which was in none of the selections for blf-classification, is

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

more remarkable. The context features seem to 'activate' utterance features in blf-classification, which is an indication of them being dependent. The feature **containsTell** on the other hand is in only one of the selections, like it was in the case of window-size 1. It is also very clear that the context features **blf-1** and **flf-1** are in all selections, which already seems plausible from the significant gain in classifier accuracies (Tables 3 and 4).

The result that the accuracies for window-size 3 were not significantly higher is partly illustrated by the fact that the additional context features **blf-2** and **flf-2** are not in all selections, where the context features for window-size 2 did appear in all selections. The fact that **blf-2** is the least informative of all context features, can be explained from the definitions of blf and flf in the annotation scheme: the backward-looking function characterises how the utterance refers to the previous dialogue, and therefore **blf-2** only refers to parts of the dialogue not taken into account in this model; however, the forward-looking function characterises how the utterance influences the future dialogue, and therefore **flf-2** refers to parts of the dialogue that *are* taken into account in this model.

The remark in Section 6.3 that the features **containsLocativePrep** and **containsLocativeAdverb** are expected to be less important in the SCHISMA dialogues than in dialogues concerning navigation and should be modified, is confirmed by the experimental results. Only a few of all optimal selections contain these features. This goes for the features **containsTell** and **startsWithCanYou** as well. These four features are therefore subject to either modification or removal in our process of finding a set of features that results in high accuracies for blf- and flf-classification.

In summary, we may draw the following tentative conclusions:

Conclusions concerning classification accuracies

1. For all obtained classifiers, the accuracies are significantly higher compared to the baseline.
2. In the 2-gram model the accuracies are significantly better than in the 1-gram model, especially in the case of blf-classification.
3. The 3-gram model does *not* yield significant improvement in accuracy compared to the 2-gram model.
4. Although not significantly, the Naive Bayes classifiers all show slightly lower accuracies compared to those of the Bayesian Network and Decision Tree classifiers.
5. The Bayesian Network and Decision Tree classifiers do not show significant differences in terms of accuracy.

Conclusions concerning feature selection

1. Based on the number of times the various features occur in the selections that yielded maximal accuracies, the features **lenq**, **startsWithWHExpr**, **endsWithQuestionMark**, **containsNegative** and **contains-Okay** are the most informative for classification.

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

2. In the same sense, **startsWithCanYou** and **containsLocativeAdverb** are the least informative features.
3. For flf-classification, more utterance features are in the optimal selections than for blf-classification.
4. Utterance features that did not seem very informative for classification, gain informativeness when context features are included, especially in the case of blf-classification.
5. The context features **blf-1** and **flf-1** are relevant in both flf- and blf-classification in the case of window-size 2.
6. In the window-size 3 model, the context features are also important; the feature **blf-2** however is the least informative.
7. The results confirm the expected irrelevancy of the features **containsLocativePrep** and **containsLocativeAdverb**.

The experiments were repeated after some modifications to the set of linguistic features. In the new model we tried out a modified version of the feature **lenq**: instead of three possible values, the feature now has four possible values: 'one' (1 word), 'few' ($2 \leq \#words < 5$), 'several' ($5 \leq \#words < 10$), and 'many' (10 or more words). The features **startsWithCanYou** and **containsTell** were removed, because the results in Section 7.1 revealed that they were not very informative features for classification. Finally, the features **containsLocativePrep** and **containsLocativeAdverb** were replaced by their 'temporal counterparts': **containsTemporalPrep** – in the specification of which prepositions such as "naar", "achter" and "langs" were replaced by prepositions such as "tot" *(Eng.: until)* and "na" *(Eng.: after)* – and **containsTemporalAdverb** – in the specification of which adverbs such as "hier", "daar" and "rechts" were replaced by adverbs such as "nu" *(Eng.: now)*, "vandaag" *(Eng.: today)* and "morgen" *(Eng.: tomorrow)*.

Unfortunately, the results for the modified set of linguistic features did not show significant improvements of the maximum accuracies. However, the new features **containsTemporalPrep** and **containsTemporalAdverb** occur more frequently in the optimal selections than their locative counterparts in the previous model. In this respect our expectation that temporal features would be more informative than locative features in the kind of dialogues we are dealing with in the corpus has been confirmed.

7.2 Alternative ngram generation

In the method of extracting the appropriate data from the corpus in the previous sections, the notion of 'previous flf/blf' is taken as 'the blf/flf of the previous turn'. In generating the data, the n-grams of segments – to which the dialogue act types have been assigned – are actually based on n-grams of turns (see Section 6.4). Because a turn may contain several segments, several sequences of n segments from a n-gram of turns are possible. We have chosen to collect all of those possible sequences for the data (this explains why the datasets for larger window-sizes are larger and the associated baselines for classification accuracy are different).

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

Table 6. *Maximum Accuracies for blf-classification in the model that considers n-grams of segments instead of turns*

Context	Naive Bayes	Bayes Net	Decision Tree	Baseline
1-grams	70.89 (±3.06)	71.12 (±3.06)	70.41 (±3.08)	64.38
2-grams	87.69 (±2.22)	87.81 (±2.21)	88.28 (±2.17)	64.38
3-grams	87.69 (±2.22)	87.93 (±2.20)	88.40 (±2.16)	64.38

Table 7. *Maximum Accuracies for flf-classification in the model that considers n-grams of segments instead of turns*

Context	Naive Bayes	Bayes Net	Decision Tree	Baseline
1-grams	62.25 (±3.27)	67.69 (±3.15)	67.57 (±3.16)	46.75
2-grams	68.88 (±3.12)	72.66 (±3.01)	73.73 (±2.97)	46.75
3-grams	70.41 (±3.08)	72.66 (±3.01)	73.73 (±2.97)	46.75

Table 8. *Maximum Accuracies for blf-classification, considering context-sizes up to and including 5, using context features only*

Context	Naive Bayes	Bayes Net	Decision Tree	Baseline
2-grams	82.96 (±2.54)	83.08 (±2.53)	84.02 (±2.47)	64.38
3-grams	83.55 (±2.50)	83.55 (±2.50)	84.02 (±2.47)	64.38
4-grams	83.55 (±2.50)	83.55 (±2.50)	84.02 (±2.47)	64.38
5-grams	83.55 (±2.50)	83.55 (±2.50)	84.02 (±2.47)	64.38

As an alternative, a corpus dialogue may be considered as one sequence of segments disregarding the turns and utterances and we may generate n-grams of segments that are actually subsequent. In each n-gram, the associated speaker will no longer be fixed, except for the last segment which is always associated with the client. This also means that the number of instances generated is the same for all window-sizes.

The results in Tables 6 and 7 show a significant improvement in comparison with the results in Section 7.1.

Some further experiments were done concerning larger context-sizes, see Tables 8 and 9. Window-sizes up to and including 5 were examined, where the linguistic features were not taken into account. Although we have not seen any improvement for context sizes of 3 and more, one interesting point can be made. In the case of blf-classification, we had a baseline of 64.38% and improved that to a maximum accuracy of 71.12% for the Bayesian Network classifier, when taking into account only linguistic features of the current segment. When taking into account only the blf and flf of the previous segment, a maximum accuracy of 83.08% is obtained for that classifier. For all 3 classifiers, adding the previous segment to the model yields over 11% higher accuracy than adding linguistic features. In the case of

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

Table 9. *Maximum Accuracies for flf-classification, considering context-sizes up to and including 5, using context features only*

Context	Naive Bayes	Bayes Net	Decision Tree	Baseline
2-grams	58.11 (±3.33)	58.11 (±3.33)	58.11 (±3.33)	46.75
3-grams	58.11 (±3.33)	60.00 (±3.30)	58.93 (±3.32)	46.75
4-grams	58.11 (±3.33)	60.00 (±3.30)	59.41 (±3.31)	46.75
5-grams	58.11 (±3.33)	60.00 (±3.30)	59.41 (±3.31)	46.75

Table 10. *Maximum Accuracies for blf-classification after adding features for making explicit who is the speaker of the segment, the client or the server*

Context	Naive Bayes	Bayes Net	Decision Tree	Baseline
1-grams	70.89 (±3.06)	71.12 (±3.06)	70.41 (±3.08)	64.38
2-grams	87.93 (±2.20)	88.28 (±2.17)	88.88 (±2.12)	64.38
3-grams	87.93 (±2.20)	88.28 (±2.17)	89.11 (±2.10)	64.38

flf-classification however, adding context instead of linguistic features produces results that are over 4% worse for the Naive Bayes classifier and over 9% worse for the Bayesian Network and Decision Tree classifier.

7.3 Further modifications

In the case of working with segments from subsequent turns, it was clear which of the speakers (i.e., client or server) was associated with a particular segment, knowing that the current segment always came from a client turn. In the alternative n-gram generation model, only the current segment is known to be a client segment, but the speakers of preceding segments were not alternately server, client, server, client, and so on, anymore. The segment directly preceding the current client segment can be another client segment. In order to take into account the speakers associated with the different segments, they need to be made explicit by means of an additional attribute.

After adding new context features that make explicit the speaker for each preceding segment, we repeated the experiments. The feature **sp-n** refers to the speaker of the n-th previous turn, so **sp-1** is the speaker of the segment directly preceding the current segment. Naturally, these new features all have three possible values: 'C' (the client), 'S' (the server) and 'null' (this segment in n-gram falls beyond the beginning of the dialogue, i.e., does not exist). The results in Tables 10 to 13 show that in the optimal accuracies, only very small improvements could be achieved in comparison to the previous model, in which the speaker in previous segments was not made explicit.

Another modification in the experiments we introduced is not in the models themselves, but in the collection of the results. In the experiments so far, we have only collected single feature selections to be optimal in the sense of yielding

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

Table 11. *Maximum Accuracies for flf-classification after adding features for making explicit who is the speaker of the segment, the client or the server*

Context	Naive Bayes	Bayes Net	Decision Tree	Baseline
1-grams	62.25 (±3.27)	67.69 (±3.15)	67.57 (±3.16)	46.75
2-grams	68.99 (±3.12)	73.02 (±2.99)	74.08 (±2.95)	46.75
3-grams	70.65 (±3.07)	73.02 (±2.99)	74.08 (±2.95)	46.75

Table 12. *Maximum Accuracies for blf-classification, with added features for indicating the speakers of previous segments, and considering larger window-sizes, without utterance features*

Context	Naive Bayes	Bayes Net	Decision Tree	Baseline
2-grams	83.67 (±2.49)	84.85 (±2.42)	85.44 (±2.38)	64.38
3-grams	83.79 (±2.49)	85.21 (±2.39)	85.80 (±2.35)	64.38
4-grams	84.02 (±2.47)	85.21 (±2.39)	85.80 (±2.35)	64.38
5-grams	84.02 (±2.47)	85.21 (±2.39)	85.80 (±2.35)	64.38

Table 13. *Maximum Accuracies for flf-classification, with added features for indicating the speakers of previous segments, and considering larger window-sizes, without utterance features*

Context	Naive Bayes	Bayes Net	Decision Tree	Baseline
2-grams	58.11 (±3.33)	58.70 (±3.32)	58.82 (±3.32)	46.75
3-grams	58.11 (±3.33)	60.59 (±3.29)	59.65 (±3.31)	46.75
4-grams	58.11 (±3.33)	60.95 (±3.29)	60.47 (±3.30)	46.75
5-grams	58.11 (±3.33)	60.95 (±3.29)	60.47 (±3.30)	46.75

maximum accuracies for a classifier. However, it may very well be the case that there are *several* selections of features that yield the same maximal classifier accuracy. In retrospect, this is not very surprising as we are using relatively small datasets and therefore even smaller datasets for testing in a cross-validation fold. Therefore, we have now modified the software to collect *all* feature selections that yield the maximum accuracy. With this in mind, we have chosen to modify the setup of the detailed result tables (see Table 14) in such a way, that each classifier is now associated with two feature selection columns: one column to indicate which features were in ALL of the optimal selections, and one column to indicate which features were in ANY of the optimal selections. Unfortunately, we are forced to leave extensive experiments in this modified approach and analyses of the results for future work.

Figure 4 shows the Bayesian network (structure) that resulted from applying the K2 algorithm on one of the optimal selections in the experiments. This network expresses various dependencies among the features, in contrast to a Naive Bayes classifier that considers all features to be independent given the

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

Table 14. *Results from blf-classification with window-size 2, taking into account multiple feature selections yielding an optimal accuracy*

Classifier	NaiveBayes		BayesNetK2		J48	
Maximum Accuracy	87.929		88.284		88.8757	
Feature Selections	All	Any	All	Any	All	Any
sp-1	+	+	+	+	+	+
flf-1	+	+	+	+	+	+
blf-1	−	−	−	+	+	+
lenq	+	+	−	+	−	−
isContinue	+	+	−	+	−	+
startsWithWHExpr	−	−	−	+	+	+
endsWithQuestionMark	−	−	−	+	+	+
startsWithCanI	−	−	−	+	−	+
startsWithIWant	+	+	−	+	+	+
containsPositive	−	−	+	+	−	−
containsNegative	+	+	+	+	+	+
containsOkay	−	−	+	+	−	−
containsTemporalPrep	−	−	−	+	−	−
containsTemporalAdverb	−	−	−	+	−	+
containsDo	−	−	−	+	−	+

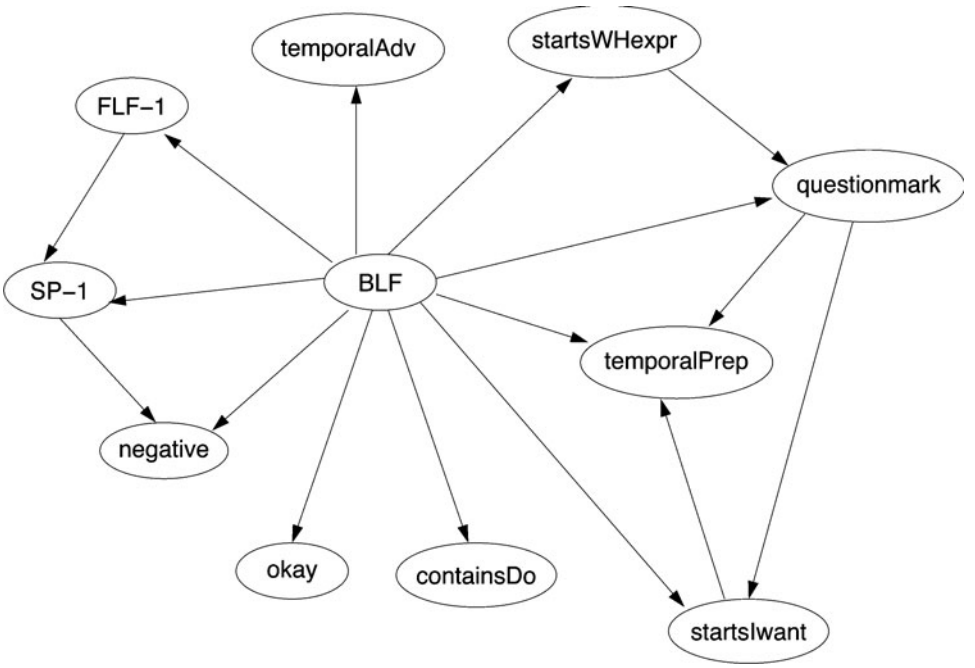


Fig. 4. Bayesian network resulting from K2 algorithm for an optimal feature selection.

class value. For example, the features **okay** and **containsDo** are assumed to be conditionally independent given the class variable **blf**, but **startsWithWHExpr** and **endsWithQuestionMark** are not.

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

8 Conclusion

In this paper we have presented work on using Bayesian networks in dialogue systems for dealing with uncertainty. With the primary aim to be able to perform some initial experiments with the specific task of dialogue act classification we developed a dialogue act annotation scheme for the SCHISMA corpus to extract training data from it. This annotation scheme is based on a standard for annotation of task-oriented dialogues, the DAMSL scheme. Our scheme contains multiple layers, including the layers of forward- and backward-looking functions. Manually annotating the corpus has resulted in 64 XML-files, each containing one annotated dialogue from the corpus.

The experiments on dialogue act classification were intended to compare Bayesian network classifiers with other classifiers and machine learning techniques. The additional aim was to find utterance and context features that are most informative for this task. The various models we have used were given by 1) a set of utterance features with possible values, 2) two sets of dialogue act types to distinguish between, i.e., one set of forward-looking functions and one set of backward-looking functions, and 3) the size of the dialogue history to be taken into account, i.e., the window-size. In each series of experiments, Bayesian network classifiers were compared with a Naive Bayes classifier and a Decision Tree in terms of classification accuracy. Furthermore, the evaluations were performed for all possible subsets of the specified features.

The results from the experiments as presented in this paper are subdivided into a number of phases that were passed through, each phase involving a modification of the models used. From the results in the initial model (Section 7.1), the most important conclusion to be drawn was that using the dialogue act type of the previous utterance significantly improved the performance of the classifiers, but taking into account further preceding utterances did not show improvement. In comparing the three classifiers evaluated, the Naive Bayes classifiers performed somewhat worse than the Bayesian network and Decision Tree.

The utterance features in the models were defined using a basic intuition regarding the relation between dialogue act types and part-of-speech information and surface patterns, including the occurrence of keywords. Although several minor conclusions regarding the significance of the utterance features could be drawn from the results concerning evaluation of subsets of this initial feature selection, the various optimal selections did not show a strong overall correspondence. Obviously, much improvement may be expected in finding appropriate features to be used in classifying dialogue act types. Nevertheless, one conclusion is that the utterance features were more relevant for classifying forward-looking functions than for backward-looking functions, whereas context features were more relevant for classifying backward-looking functions than for forward-looking functions. This result confirms our prior intuition regarding the definition of forward- and backward-looking functions.

The modification of the classification models as presented in Section 7.2 resulted in significantly improved accuracies. The modification involved a change in the sense of context, i.e., instead of looking at sequences of turns the new model concerns sequences of segments. As the segments considered in the new model were not

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

alternately associated with client and server, the speaker of a segment was made explicit by an additional feature, a modification introduced in Section 7.3.

Another modification described in Section 7.3 concerns the analysis of subsets of features in the classification experiments. This modification involves an improved account of the fact that in several cases, multiple subsets of features resulted in the maximal classifier accuracy, instead of the single first subset recorded and analysed in the previous models.

8.1 Future recommendations

Future work will have to involve specifying additional utterance features and modifying and/or removing existing ones, in order to obtain higher dialogue act classification accuracies. The motivation behind the set of linguistic features is quite intuitive and is not accounted for in a very systematic and elaborate way. In this light, further attention to pre-selecting linguistic features seems worthwhile in order to attain higher classification accuracies.

Inspection of more detailed information about the classification results might be useful, for example, the results per class value as stored in a confusion matrix and in the form of precision/recall results.

If we are able to obtain better classification results, we might also be better able to pick one set of the most significant features to do further experiments with. These further experiments will be aimed more at the comparison of different classifiers with different parameter sets.

Another suggestion for improving classification results might be the specification of two different sets of dialogue acts for client and server utterances. In classifying client utterances we might not be interested in distinguishing between two dialogue act types, say, A and B, but this distinction might be relevant for any (previous) server utterances taken into account in the classification. The differences in frequencies of the dialogue act types for client and server in the corpus used also suggest this (Tables 1 and 2).

In the conclusions concerning window-sizes it was noted that the inclusion of one preceding utterance in the model (a window-size of 2) yielded a significant improvement in the classification accuracies, but including further preceding utterances did not. Experiments with models including context features only, where window-sizes up to 5 were explored, suggested that looking back any further in the dialogue seems useless, because no improved accuracies were found beyond window-size 2. However, one factor that could give rise to improved accuracies for larger window-sizes, has not been taken into account. If the dialogues are deeply structured with sub-dialogues, e.g., for verification or repair, or sub-dialogues related to the task at hand, then it may occur that the current utterance is produced after completion of a sub-dialogue and elaborates on the utterance preceding that sub-dialogue. Therefore, the size of possible sub-dialogues should be taken into account when deciding on the window-sizes to be explored in the classification experiments.

Finally, we would like to mention the option of extending the annotation scheme with more specific dialogue act types and extend the annotated corpus accordingly.

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

This may result in data leading to higher classification accuracies, but only for this specific type of dialogue. This reveals a trade-off in developing dialogue act classifiers. On the one hand, specific dialogue acts can be defined for specific types of dialogue, where the classification task is relatively simple, but the resulting classifiers may not work well for other types of dialogue. On the other hand, task-independent, general dialogue act types can be used, where the classification task is more complex and more data will be needed, but the resulting classifiers perform equally well in various types of dialogue.

An alternative way to exploring possible sets of dialogue act types is by unsupervised learning of dialogue act classes, based on superficial utterance features. These classes can be used in a classifier instead of the (subjectively) selected dialogue act types. Experiments – using the same corpus – with this approach have been done for two different unsupervised learning techniques: the AutoClass program (Andernach, 1996) and Kohonen Networks (Andernach *et al*., 1997). Similar work has been done by (Jokinen *et al*., 2001), who have used self-organising maps (SOM) and LVQ-classification in their experiments on a small dialogue corpus.

References

Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Kipp, M., Koch, S., Maier, E., Reithinger, N., Schmitz, B. and Siegel, M. (1998) Dialogue acts in VERBMOBIL-2 second edition. VM-Report 226, DFKI GmbH.

Allen, J. and Core, M. (1997) Draft of DAMSL: Dialog Act Markup in Several Layers. Dagstuhl Workshop.

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

Andernach, T., Poel, M. and Salomons, E. (1997) Finding classes of dialogue utterances with Kohonen networks. In: Daelemans, W., van den Bosch, A. and Weijters, A., editors, *Workshop Notes of the ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks*, pp. 85–94, Prague, Czech Republic.

Austin, J. (1975) *How to Do Things with Words*. Harvard University Press, 2nd edition.

Bilvi, M. and Pelachaud, C. (2003) Communicative and statistical eye gaze predictions. *Workshop Embodied Conversational Characters as Individuals (in conjunction with AAMAS 2003*.

Black, W., Thompson, P., Funk, A. and Conroy, A. (2003) Learning to classify utterances in a task-oriented dialogue. In: Jokinen, K., Gambäck, B., Black, W., Catizone, R. and Wilks, Y., editors, *Proceedings of the Workshop on Dialogue Systems: Interaction, Adaptation and Styles of Management*, pp. 9–16, Budapest.

Bunt, H. (1995). Dynamic interpretation and dialogue theory. In: Taylor, M., Bouwhuis, D. G. and Neel, F., editors, *The Structure of Multimodal Dialogue*, volume 2. John Benjamins, Amsterdam.

Bunt, H. (1996) Interaction management functions and context representation requirements. In: Luperfoy, S., Nijholt, A. and van Zanten, G. V., editors, *Dialogue Management in Natural Language Systems*, pp. 187–198. University of Twente.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G. and Anderson, A. (1996). HCRC Dialogue Structure Coding Manual. Technical Report TR-82, HCRC, Edinburgh.

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

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

Galley, M., McKeown, K., Hirschberg, J. and Shriberg, E. (2004) Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pp. 669–676, Barcelona, Spain.

Goffman, E. (1981) *Forms of Talk*, chapter Replies and Responses, pages 5–77. University of Pennsylvania Press.

Heckerman, D. and Horvitz, E. (1998) Inferring informational goals from free-text queries: A Bayesian approach. *Fourteenth Conference on Uncertainty in Artificial Intelligence*, pp. 230–237, Madison, WI. Morgan Kaufmann.

Hofs, D., op den Akker, R., Nijholt, A. and Hondorp, H. (2003) A multimodal interaction system for navigation. In: Kruijff-Korbayova, I. and Kosny, C., editors, *Proceedings Diabrück 2003, 7th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 183–184, Saarbrücken, Germany.

Jokinen, K., Hurtig, T., Hynnä, K., Kanto, K., Kaipainen, M. and Kerminen, A. (2001) Self-organizing dialogue management. *Proceedings of the NLPRS Workshop 'Neural Networks and Natural Language Processing'*, pp. 78–85, Tokyo, Japan.

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

Keizer, S. (2001b) Dialogue act modelling using Bayesian networks. In: Striegnitz, K., editor, *Proceedings of the Sixth ESSLLI Student Session*, pp. 143–153, Helsinki.

Keizer, S. (2003) *Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks*. Number 2003–11 in SIKS Dissertation Series. Twente University Press, Enschede, The Netherlands.

Keizer, S., op den Akker, R. and Nijholt, A. (2002) Dialogue act recognition with Bayesian networks for Dutch dialogues. *Proceedings 3rd ACL/SIGdial Workshop on Discourse and Dialogue*, pp. 88–94, Philadelphia, PA.

Kipp, M. (1998) The neural path to dialogue acts. In: Prade, H., editor, *Proceedings of the 13th European Conference on Artificial Intelligence*, pp. 175–179.

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

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.

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

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

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

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

Poesio, M. and Traum, D. (1998) Towards an Axiomatization of Dialogue Acts. In: Hulstijn, J. and Nijholt, A., editors, *TwenDial'98: Formal Semantics and Pragmatics of Dialogue*, no. 13 in TWLT.

Pulman, S. (1996) Conversational games, belief revision and Bayesian networks. In: Landsbergen, J., Odijk, J., van Deemter, K. and van Zanten, G. V., editors, *Computational Linguistics in the Netherlands*. SRI Technical Report CRC-071.

Reithinger, N. and Klesen, M. (1997) Dialogue act classification using language models. In: Kokkinakis, G., Fakotakis, N. and Dermatas, E., editors, *Proceedings of the 5th European*

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press

Conference on Speech Communication and Technology EuroSpeech-97, volume 4, pp. 2235–2238, Rhodes.

Samuel, K., Carberry, S., and Vijay-Shanker, K. (1998). Dialogue act tagging with transformation-based learning. *Proceedings of the 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics (ACL-COLING)*, pp. 1150–1156.

Searle, J. (1969) *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Schegloff, E. and Sacks, H. (1973) Opening up closings. *Semiotica*, **8**:289–327.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J. and Carvey, H. (2004) The ICSI meeting recorder dialog act (MRDA) corpus. In: M. Strube and C. Sidner, editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pp. 97–100, Cambridge, Massachusetts, April 30–May 1 2004. Association for Computational Linguistics.

Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Ess-Dykema, C. V., Ries, K., Shriberg, E., Jurafsky, D., Martin, R. and Meteer, M. (2000) Dialogue act modelling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, **26**(3): 339–374.

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

Witten, I. and Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann. URL: http://www.cs.waikato.ac.nz/~ml/weka/.

Wittgenstein, L. (1958) *Philosophical Investigations*. Blackwell Publishers, 2nd edition.

Wright, H. (1998) Automatic utterance type detection using suprasegmental features. *Int. Conf. on Spoken Language Processing*, volume 4, pp. 1403–1406, Sydney.

Wright, H., Poesio, M. and Isard, S. (1999) Using high level dialogue information for dialogue act recognition using prosodic features. *ESCA Tutorial and Research Workshop on Dialogue and Prosody*, pp. 139–143.

https://doi.org/10.1017/S1351324905004067 Published online by Cambridge University Press