MODELLING MULTIMODAL USER ID IN DIALOGUE

Hartwig Holzapfel and Alex Waibel

interACT Research
Interactive Systems Labs
Universität Karlsruhe (TH), Germany
{hartwig,waibel}@ira.uka.de

ABSTRACT

This paper presents an approach to model user ID in dialogue. A belief network is used to integrate ID classifiers, such as face ID and voice ID, and person related information, such as the first name and last name of a person from speech recognition or spelling. Different network structures are analyzed and compared with each other and are compared with a rule-based user model. The approach is evaluated on dialogue data collected in a person identification scenario, which includes both, identification of known persons and interactive learning of names and ID of unknown persons.

1. INTRODUCTION

Modeling user ID in dialogue is of great interest for different systems. For example in human-robot interaction the robot can distinguish different persons only when equipped with a reliable user (ID) model. Modelling the user's ID is a typical classification problem, given a set of input parameters and a classification result. A typical affordance in an interactive dialogue system is that hypotheses are computed during runtime of the system, i.e. while the dialogue continues, instead of collecting all relevant information and only then applying classification. In fact, the dialogue flow is also influenced by the decisions that are made by the ID classifier, leading to better results and shorter dialogues, if the model can generate good hypotheses early in the dialogue. The approach proposed here has been examined for human-robot interaction, where the users engage in explicit identification dialogues. For example, the robot can ask a user for the name, or use explicit or implicit confirmation strategies given different kind of observations. Perceptual technology used for the experiments is based on sensors typically used on a humanoid robot, such as stereo vision and speech recognition. For simplicity we frequently use the term user model in the following, which in this paper refers to modelling the user's ID.

In this work we propose a user model that combines information collected during dialogue, such as spoken names, spelled names, confirmations and multimodal ID classification from face ID and voice ID. Fusion of these modalities is done using Bayesian (belief) networks. Bayesian networks are adequate for this approach, since they model probabilities of observations given a true state of nature. A key aspect is estimating conditional probabilities, such as confidence measures for multimodal ID hypotheses. Generally speaking, these confidence measures are necessary to cope with recognition errors. For example, if a first name has been misrecognized and contradicts the multimodal ID hypothesis, the system computes the best hypothesis while taking into account probabilities of misrecognition of each input hypothesis. It can then ignore the incorrect speech recognition if multimodal ID confidence is high enough.

Bayesian networks are frequently used in data mining, to discover statistical dependencies on large data sets, with the goal of learning network structures. In our work, the network structure is created manually and different network structures are analyzed. The following chapter describes some basic properties, a more detailed overview can be found for example in [1]. In [2] a belief network has been used for multimodal user registration. It is similar to the 'simple' network structure presented in the following and compared to more complex structures.

In contrast to other work our approach takes into account unknown persons, unknown word detection plus name spelling, and features extracted from the dialogue history. Special attention is given to confidence estimates. The presented approach can generally be extended with other features. For example on could consider day of time when a person interacts with the system and integrate this as a conditional probability. The approach also shows significant improvement over our previous identification model [3].

2. BELIEF NETWORKS FOR PERSON IDENTIFICATION

A Bayesian network is a directed acyclic graph with nodes and edges. Each node represents a variable which is either discrete or continuous, and edges are modeled as conditional probabilities. Depending on the type of variables the network is either a discrete, continuous, or a hybrid bayesian

network. In the presented work, we use a discrete network. When some variables are observed (they are then called evidence variables) other variables in the network can be queried using probabilistic inference.

2.1. Input Features and Network Structure

In our network we use three categories of observations as evidence for identification. Evidence corresponds to information slots filled by the dialogue system. The first observation category is multimodal ID (MMID) classification which directly classifies the person's ID. The second type of observation only provides hints about the person's ID but doesn't classify one person exclusively. Such observation is recognition of the spoken first name. In our model, a person can have only one first name; however, different persons (either known or unknown) may share the same first name. The third type of observation is extracted from the dialogue flow, such as disconfirmed names.

These types of observation can be modeled in the belief network as the following discussion shows. The structure of the belief network is determined by the definition of conditional probabilities. A standard ID classifier produces hypotheses with posterior probabilities, i.e. $P(ID|observation)$. Another way of modeling, which also describes a causal structure, is the inverted dependency structure $P(classification\text{-}correct|ID)$ with a directed edge from 'ID' to 'classification-correct'. Now, the conditional probability models the probability of a classification being correct given the ID. This probability is estimated by independent confidence classification which is multiplied by the n-best list hypothesis score with successive normalization. Confidence estimation for multimodal ID is described in section 2.3. Using Bayes theory to combine different classification results leads to some practical issues with the extreme values 0 and 1. This is the case, e.g. when IDs are not represented in the n-best list, thus additional factors and an offset are introduced with the following formula:

$$w_{id} = m + conf * a(2 * score_{id} - 1) \quad (1)$$

The values 'm' (offset), the confidence of the classifier and the scaling factor 'a' influence the rating of the original ID-score from the hypothesis list. The desired probability is then obtained by normalizing w_{id} by the sum over all w_{id}.

In a similar way, spoken name recognition (speech recognition results) is integrated into the network as evidence. A conditional probability $P(name\text{-}correct|name)$ models first name and last name recognition. An additional edge $P(name|ID)$ connects names to IDs. It is set to 1 for persons that have been entered manually and can be set to a smaller value to model uncertainty in the knowledge base when a person has been learned interactively.

A fourth type of information is not used as evidence in the network, but influences conditional probabilities in the network. For example confirmation of a name is a feature that is observed by the dialogue model. In this case the evidence, i.e. the value of the observed name, doesn't change, but the probability of the the name being correct increases.

2.2. Network Structure

Figure 1 shows the structure of a simple belief network integrating multimodal ID, first name and last name recognition. An abstract ID node represents the ID of a person; other nodes represent evidence as pointed out above.

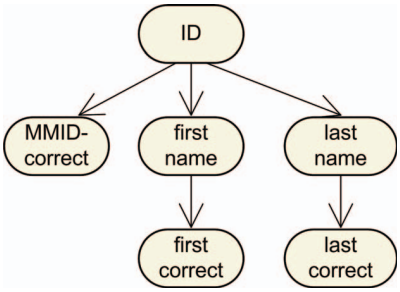

Fig. 1. Simple user ID belief network

The simple network structure works well for many situations with only known persons. However, some important aspects are missing. For example the network doesn't model the reduced probability of a name after it has been rejected. For rejected (disconfirmed) names a separate blacklist is added. It accounts for the fact that also rejections are error prone. A name on the blacklist is assigned 1/10 of the probability of non-rejected names. Also the problem of unknown first name / last name combinations is addressed. An unknown detection node increases the likelihood of an unknown person by 100 * prior user probability if first name / last name combinations are observed that don't match the database of known persons. The factor 100 has been chosen experimentally to 'compete' against multimodal user ID. Figure 2 shows the extended network structure.

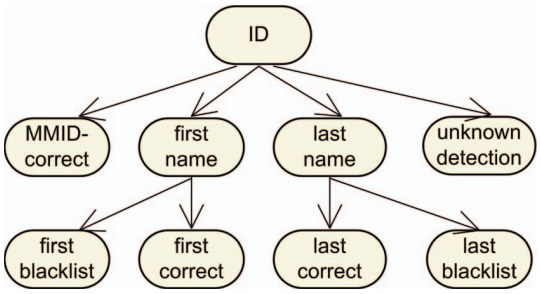

Fig. 2. Extended user ID belief network with blacklist and unknown model

Some considerations had to be made so that the proposed user ID model can be used in an online system. The main con-

siderations relate to dynamic updates in the network. Some parts of the network are static, which are the structure of the network and the node names. Node values, i.e. person IDs, first names, last names, etc. are generated automatically from database entries. Edges in the network, i.e. conditional properties, are also updated dynamically, e.g. the edge representing correct recognition of multimodal ID is updated with each new hypothesis according to the confidence value.

2.3. Multimodal ID and Confidence Measures

As mentioned before, confidences are very helpful in theory to estimate a 'trust-level' of a classifier, especially when hypotheses from different classifiers are combined. In [4] we have proposed an approach for confidence-based fusion of face ID and voice ID, which uses logistic regression for confidence estimation on several levels (on single hypothesis, sequence hypothesis and output confidence). We use this approach to model the multimodal ID ('MMID') node in the belief network, but restrict ourselves to using video information, leaving out voice segments. The reason for this is that even though the recognition rates are better with voice information, no sufficient voice data was available for independent training and evaluation. The belief network uses the hypothesis score (from n-best list) plus the classification confidences, as described in the beginning of this section.

In the experiments reported here, this approach has been used in two configurations. The first configuration is closed-set person identification, where the MMID classifier always decides on a label known from training data. The second configuration is open-set person identification, where there is an additional category 'unknown' to classify persons which are not in the training set. To integrate the unknown classification in the n-best list, we estimate the hypothesis score by 1.0 minus classifier confidence, which produces stable results on the given corpus.

3. EXPERIMENTS

3.1. Dialog Data Collection

Data used for experiments has been collected during different robot receptionist dialogs with user ID and name learning [3]. The dialogue manager uses different strategies (a fixed strategy was employed per scenario) to identify a person's ID, first name and last name. During the dialogues, speech and image sequences have been recorded for voice ID and face ID, speech recognition results have been logged and all interactions have been transcribed. From this data we obtain a corpus of annotated sessions, with a timeline of events including all dialogue system input.

With this data we have then conducted the evaluation of different user models. The advantage of the approach is that once data has been collected, different user models can be compared on the same data. While there is an effect of the applied user model on the dialogue flow, recorded information can be observed by all models. Thus, a user model that has not been used for recording will be slightly underestimated. The best comparison can be drawn by the end of an evaluated dialogue session.

3.2. Baseline Approach

The baseline or 'confirmation' approach uses a rule-based system and a confirmation strategy to determine the ID. It uses three slots: *MMID*, *firstName*, *lastName* with different slot states, and the output slot *userID*. *userID* and *MMID* have the states EMPTY, SET, CONFIRMED. *firstName* and *lastName* have the states EMPTY, UNKNOWN, SPOKEN, SPELLED, CONFIRMED. A slot is set to EMPTY when the information slot is empty or when its value has been disconfirmed. The update rule for setting the *userID* value takes into account reliability of the slot values. For example CONFIRMED has the highest reliability, and spoken name input is preferred over multimodal ID. The latter only is considered when first name and last name slots are empty, which happens typically at the beginning of a dialogue or after a name has been rejected.

3.3. Evaluation

The evaluation compares different configurations of the bayes-network with each other and against the confirmation approach. The evaluation has been conducted on two conditions: person is known (i.e. has talked to the robot before and MMID training data is available) vs. person unknown (no training data available). Each condition is evaluated with open set vs. closed set MMID classification, each with a database of 25 known persons. In the unknown condition, 46 sessions are available for evaluation; in the known condition 43 sessions are available. Before evaluation, MMID models have been trained with independent training data for the two conditions: known and unknown. The set for the unknown condition includes all sessions from the known condition, however the respective person was excluded from the MMID training data and the person database.

To evaluate the approach, different metrics are used. 'UID rows' is the percentage of all correct hypotheses, i.e. all input event during the interaction. 'UID end' is the percentage of all correct final hypotheses, i.e. the last hypothesis of each dialogue. 'UID norm' is a normalized correct rate, i.e. the average correct rate per dialogue. It prevents that long sessions get higher weight than short sessions. For example, the shortest sessions without face ID input has only 6 input events, in contrast to the longest session with 250 input events.

Table 1 reports recognition rates of the multimodal ID classifier representing the observations of the MMID node. Table 2 shows the numbers from the evaluation runs with closed set and open set classification. The user models listed in the tables are the baseline 'confirm' model, the simple

set /condition	events	MMID	sessions	MMID end
closed / known	1870	84,44%	43	81,4%
closed / unk	2165	0,00%	46	0,0%
open / known	1870	80,53%	43	74,4%
open / unk	2165	89,01%	46	89,1%

Table 1. Task overview: number of input events, MMID per event, number of sessions, MMID at end of the session.

bayes model 'bayes-p' without black list, the 'bayes-bnr' model with black list but without resetting of user names after disconfirm, the 'bayes-blu' model including black list and unknown person detection, and the 'bayes-bl' model with black list and resetting of names. The unknown/closed set has been excluded since models are not suitable for this category, only bayes-blu and bayes-bl achieve 100% for UID end, the others achieve 0.0%.

The overall best model is the 'bayes-bl' model, which outperforms the 'bayes-blu' model in the open-set condition, where the unknown detection from face ID is more reliable than unknown detection from name recognition and static properties of the bayes-bl network. In the closed-set condition both are almost equal. In the closed set/known condition the simple model obviously performs best, since it doesn't produce false alarms for unknown.

The 'UID-norm' value looks worse than 'UID rows' for most conditions. This is reasonable since some sessions don't have any face ID at all. These sessions start without relevant information and only at the end of a session a good hypothesis can be found by the model. In general this also mirrors the fact that user ID hypothesis improves with the dialogue flow.

Given the kind of evaluation with a static dialogue corpus, the effect of the user model on the dialogue flow cannot be measured. Despite the fact that the dialogues had been recorded with the baseline user model, the results show that the belief network operates more reliably than the baseline model. The recognition rates are better especially at the end of the dialogue. This significantly improves the robot's perception who the robot is talking to and improves memorizing persons. For a detailed analysis, how the model effects the dialogue flow, additional experiments need to be conducted.

4. CONCLUSIONS

In this work we have presented an approach for multimodal integration to model user ID in an interactive dialogue system. The approach considers aspects of an online system where information is delivered and updated sequentially. The approach also considers special aspects of a dialogue system where information is confirmed or rejected during dialogue.

We have compared different belief network structures with each other and against a baseline model that purely relies on dialogue information with confirmation and rejection

condition	task	UID rows	UID end	UID norm
known/c	confirm	85.94%	83.7%	74.77%
known/c	bayes-p	83.74%	95.4%	77.90%
known/c	bayes-bnr	75.13%	93.0%	76.52%
known/c	bayes-blu	73.32%	93.0%	75.13%
known/c	bayes-bl	79.68%	93.0%	76.41%
known/o	confirm	77.38	72.1%	67.19%
known/o	bayes-p	80.80	95.4%	73.54%
known/o	bayes-bnr	72.19	93.0%	72.16%
known/o	bayes-blu	69.84	90.7%	70.59%
known/o	bayes-bl	76.58	93.0%	71.94%
unk/o	confirm	91.45%	100.0%	88.20%
unk/o	bayes-p	86.00%	58.7%	83.53%
unk/o	bayes-bnr	86.33%	60.9%	83.59%
unk/o	bayes-blu	91.50%	100.0%	88.11%
unk/o	bayes-bl	91.50%	100.0%	88.11%

Table 2. User ID evaluation with closed set '/c' and open set '/o' multimodal ID

of the best hypothesis. The best configurations perform better than the baseline model and are suitable for person identification in dialogue. The presented approach is also suitable for open set person identification. We think that the user model can show beneficial for a statistical dialogue model. In the future, detailed effects on the dialogue flow can be shown with experiments using the new user models.

5. ACKNOWLEDGEMENTS

This work was supported in part by the German Research Foundation (DFG) as part of the Collaborative Research Center 588 "Humanoid Robots - Learning and Cooperating Multimodal Robots". The authors would like to thank Philipp Hüthwohl who has contributed to this work as part of his student work and Philipp Grosse for multimodal ID training.

6. REFERENCES

[1] David Heckerman, "A tutorial on learning with bayesian networks," Msr-tr-95-06, Microsoft Research, 1996.

[2] Fei Huang, Jie Yang, and Alex Waibel, "Dialogue management for multimodal user registration," in *Proceedings of the International Conference for Speech and Language Processing (ICSLP)*, 2000.

[3] Hartwig Holzapfel and Alex Waibel, "Behavior models for learning and receptionist dialogs," in *Interspeech 2007*, Antwerp, Belgium, 2007.

[4] Philipp Grosse, Hartwig Holzapfel, and Alex Waibel, "Confidence based multimodal fusion for person identification," in *Proceedings of ACM Multimedia*, 2008.