Impact of ASR N-Best Information on Bayesian Dialogue Act Recognition

Heriberto Cuayáhuitl, Nina Dethlefs, Helen Hastie, Oliver Lemon
School of Mathematical and Computer Sciences,
Heriot-Watt University, Edinburgh, UK
{h.cuayahuitl,n.s.dethlefs,h.hastie,o.lemon}@hw.ac.uk

Abstract

A challenge in dialogue act recognition is the mapping from noisy user inputs to dialogue acts. In this paper we describe an approach for re-ranking dialogue act hypotheses based on Bayesian classifiers that incorporate dialogue history and Automatic Speech Recognition (ASR) N-best information. We report results based on the Let's Go dialogue corpora that show (1) that including ASR N-best information results in improved dialogue act recognition performance (+7% accuracy), and (2) that competitive results can be obtained from as early as the first system dialogue act, reducing the need to wait for subsequent system dialogue acts.

1 Introduction

The primary challenge of a Dialogue Act Recogniser (DAR) is to find the correct mapping between a noisy user input and its true dialogue act. In standard "slot-filling" dialogue systems a dialogue act is generally represented as *DialogueActType(attribute-value pairs)*, see Section 3. While a substantial body of research has investigated different types of models and methods for dialogue act recognition in spoken dialogue systems (see Section 2), here we focus on re-ranking the outputs of an existing DAR for evaluation purposes. In practice the re-ranker should be part of the DAR itself. We propose to use multiple Bayesian classifiers to re-rank an initial set of dialogue act hypotheses based on information from the dialogue history as well as ASR N-best lists. In particular the latter type of information helps us to learn mappings between dialogue acts and common mis-recognitions. We present experimental results based on the Let's Go dialogue corpora which indicate that re-ranking hypotheses using ASR N-best information can lead to improved recognition. In addition, we compare the recognition accuracy over time and find that high accuracy can be obtained with as little context as one system dialogue act, so that there is often no need to take a larger context into account.

2 Related Work

Approaches to dialogue act recognition from spoken input have explored a wide range of methods. (Stolcke et al., 2000) use HMMs for dialogue modelling, where sequences of observations correspond to sequences of dialogue act types. They also explore the performance with decision trees and neural networks and report their highest accuracy at 65% on the Switchboard corpus. (Zimmermann et al., 2005) also use HMMs in a joint segmentation and classification model. (Grau et al., 2004) use a combination of Naive Bayes and n-grams with different smoothing methods. Their best models achieve an accuracy of 66% on English Switchboard data and 89% on a Spanish corpus. (Sridhar et al., 2009; Wright et al., 1999) both use a maximum entropy classifier with n-grams to classify dialogue acts using prosodic features. (Sridhar et al., 2009) report an accuracy of up to 74% on Switchboard data and (Wright et al., 1999) report an accuracy of 69% on the DCIEM Maptask Corpus. (Bohus and Rudnicky, 2006) maintain an N-best list of slot values using logistic regression. (Surendran and Levow, 2006) use a combination of linear support vector machines (SVMs) and HMMs. They report an accuracy of 65.5% on the HCRC MapTask corpus and conclude that SVMs are well suited for sparse text and dense acoustic features. (Gambäck et al., 2011) use SVMs within an active learning framework. They show that while passive learning achieves an accuracy of 77.8% on Switchboard data, the active learner achieves up to 80.7%. (Henderson et al., 2012) use SVMs for dialogue act recognition from ASR word confusion networks.

Proceedings of the SIGDIAL 2013 Conference, pages 314–318,
Metz, France, 22-24 August 2013.

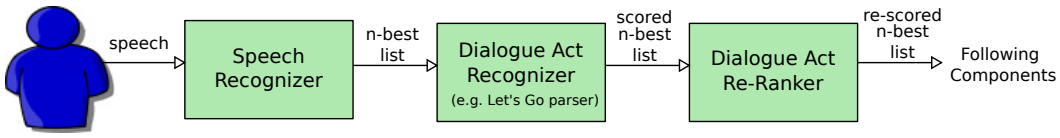

Figure 1: *Pipeline architecture for dialogue act recognition and re-ranking component. Here, the input is a list of dialogue acts with confidence scores, and the output is the same list of dialogue acts but with recomputed confidence scores. A dialogue act is represented as DialogueActType(attribute-value pairs).*

Several authors have presented evidence in favour of Bayesian methods. (Keizer and op den Akker, 2007) have shown that Bayesian DARs can outperform baseline classifiers such as decision trees. More generally, (Ng and Jordan, 2001) show that generative classifiers (e.g. Naive Bayes) reach their asymptotic error faster than discriminative ones. As a consequence, generative classifiers are less data intensive than discriminative ones.

In addition, several authors have investigated dialogue belief tracking. While our approach is related to belief tracking, we focus here on spoken language understanding under uncertainty rather than estimating user goals. (Williams, 2007; Thomson et al., 2008) use approximate inference to improve the scalability of Bayes nets for belief tracking and (Lison, 2012) presents work on improving their scalability through abstraction. (Mehta et al., 2010) model user intentions through the use of probabilistic ontology trees.

Bayes nets have also been applied to other dialogue-related tasks, such as surface realisation within dialogue (Dethlefs and Cuayáhuitl, 2011) or multi-modal dialogue act recognition (Cuayáhuitl and Kruijff-Korbayová, 2011). In the following, we will explore a dialogue act recognition technique based on multiple Bayesian classifiers and show that re-ranking with ASR N-best information can improve recognition performance.

3 Re-Ranking Dialogue Acts Using Multiple Bayesian Networks

Figure 1 shows an illustration of our dialogue act re-ranker within a pipeline architecture. Here, processing begins with the user's speech being interpreted by a speech recogniser, which produces a first N-best list of hypotheses. These hypotheses are subsequently passed on and interpreted by a dialogue act recogniser, which in our case is represented by the Let's Go parser. The parser produces a first set of dialogue act hypotheses, based on which our re-ranker becomes active. A full dialogue act in our scenario consists of three elements: dialogue act types, attributes (or slots), and slot values. An example dialogue act is *inform(from=Pittsburgh Downtown)*. The dialogue act re-ranker thus receives a list of hypotheses in the specified form (triples) from its preceding module (a DAR or in our case the Let's Go parser) and its task is to generate confidence scores that approximate true label (i.e. the dialogue act really spoken by a user) as closely as possible.

We address this task by using multiple Bayesian classifiers: one for classifying a dialogue act type, one for classifying a set of slots, and the rest for classifying slot values. The use of multiple classifiers is beneficial for scalability purposes; for example, assuming 10 dialogue act types, 10 slots, 10 values per slot, and no other dialogue context results in a joint distribution of 10^{11} parameters. Since a typical dialogue system is required to model even larger joint distributions, our adopted approach is to factorize them into multiple independent Bayesian networks (with combined outputs). A multiple classifier system is a powerful solution to complex classification problems involving a large set of inputs and outputs. This approach not only decreases training time but has also been shown to increase the performance of classification (Tax et al., 2000).

A Bayesian Network (BN) models a joint probability distribution over a set of random variables and their dependencies, see (Bishop, 2006) for an introduction to BNs. Our motivation for using multiple BNs is to incorporate a fairly rich dialogue context in terms of what the system and user said at lexical and semantic levels. In contrast, using a single BN for all slots with rich dialogue context faces scalability issues, especially for slots with large numbers of domain values, and is therefore not an attractive option. We denote our set of Bayesian classifiers as $\lambda = \{\lambda^{dat}, \lambda^{att}, ..., \lambda^{val(i)}\}$, where BN λ^{dat} is used to rank dialogue act types, BN λ^{att} is used to rank attributes, and the other BNs ($\lambda^{val(i)}$) are used to

rank values for each slot i. The score of a user dialogue act ($< d, a, v >$) is computed as:

$$P(d, a, v) = \frac{1}{Z} \prod P(d|pa_d)P(a|pa_a)P(v|pa_v),$$

where d is a dialogue act type, a is an attribute (or slot), v is a slot value, pa_x is a parent random variable, and Z is a normalising constant. This implies that the score of a dialogue act is the product of probabilities of dialogue act type and slot-value pairs. For dialogue acts including multiple slot-value pairs, the product above can be extended accordingly. The best and highest ranked hypothesis (from space $\mathcal{H}$) can be obtained according to:

$$< d, a, v >^* = \arg \max_{<d,a,v> \in \mathcal{H}} P(d, a, v).$$

In the following, we describe our experimental setting. Here, the structure and parameters of our classifiers will be estimated from a corpus of spoken dialogues, and we will use the equations above for re-ranking user dialogue acts. Finally, we report results comparing Bayesian classifiers that make use of ASR N-best information and dialogue context against Bayesian classifiers that make predictions based on the dialogue context alone.

4 Experiments and Results

4.1 Data

Our experiments are based on the Let's Go corpus (Raux et al., 2005). Let's Go contains recorded interactions between a spoken dialogue system and human users who make enquiries about the bus schedule in Pittsburgh. Dialogues are driven by system-initiative and query the user sequentially for five slots: *an optional bus route, a departure place, a destination, a desired travel date,* and *a desired travel time*. Each slot needs to be explicitly (or implicity) confirmed by the user. Our analyses are based on a subset of this data set containing 779 dialogues with 7275 turns, collected in the Summer of 2010. From these dialogues, we used 70% for training our classifiers and the rest for testing (with 100 random splits). Briefly, this data set contains 12 system dialogue act types[1], 11 user dialogue act types[2], and 5 main slots with variations[3]. The number of slot values ranges between

[1]ack, cant help, example, expl_conf, go back, hello, impl_conf, more buses, request, restart, schedule, sorry.

[2]affirm, bye, go back, inform, negate, next bus, prevbus, repeat, restart, silence, tellchoices.

[3]date.absday, date.abmonth, date.day, date.relweek, from, route, time.ampm, time.arriveleave, time.hour, time.minute, time.rel, to.

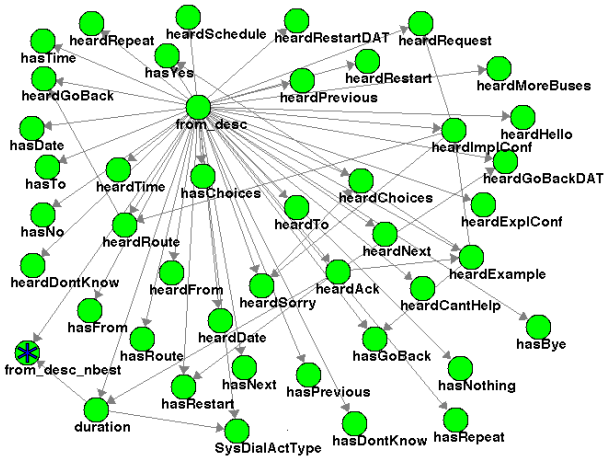

Figure 2: *Bayesian network for probabilistic reasoning of locations (variable "from_desc"), which incorporates ASR N-best information in the variable"from_desc_nbest" and dialogue history information in the remaining random variables.*

10^2 and 10^3 so that the combination of all possible dialogue act types, attributes and values leads to large amounts of triplets. While the majority of user inputs contain one user dialogue act, the average number of system dialogue acts per turn is 4.2. Note that for the user dialogue act types, we also model *silence* explicitly. This is often not considered in dialogue act recognisers: since the ASR will always try to recognise something out of any input (even background noise), typical dialogue act recognisers will then try to map the ASR output onto a semantic interpretation.

4.2 Bayesian Networks

We trained our Bayesian networks in a supervised learning manner and used 43 discrete features (or random variables) plus a class label (also discrete). The feature set is described by three main subsets: **25** system-utterance-level binary features[4] derived from the system dialogue act(s) in the last turn; **17** user-utterance-level binary features[5] derived from (a) what the user heard prior to the current turn, or (b) what keywords the system recognised in its

[4]System utterance features: heardAck, heardCantHelp, heardExample, heardExplConf, heardGoBackDAT, heardHello, heardImplConf, heardMoreBuses, heardRequest, heardRestartDAT, heardSchedule, heardSorry, heardDate, heardFrom, heardRoute, heardTime, heardTo, heardNext, heardPrevious, heardGoBack, heardChoices, heardRestart, heardRepeat, heardDontKnow, lastSystemDialActType.

[5]User utterance features: hasRoute, hasFrom, hasTo, hasDate, hasTime, hasYes, hasNo, hasNext, hasPrevious, hasGoBack, hasChoices, hasRestart, hasRepeat, hasDontKnow, hasBye, hasNothing, duration in secs. (values=0,1,2,3,4,>5).

list of speech recognition hypotheses; and **1** word-level non-binary feature (*_nbest) corresponding to the slot values in the ASR N-best lists.

Figure 2 shows the Bayes net corresponding to the classifier used to rank location names. The random variable *from_desc* is the class label, the random variable *from_desc_nbest* (marked with an asterisk) incorporates slot values from the ASR N-best lists, and the remaining variables model dialogue history context. The structure of our Bayesian classifiers were derived from the K2 algorithm[6], and their parameters were derived from maximum likelihood estimation. In addition, we performed probabilistic inference using the Junction tree algorithm[7]. Based on these data and tools, we trained 14 Bayesian classifiers: one for scoring dialogue act types, one for scoring attributes (slots), and the rest for scoring slot values.

4.3 Experimental Results

We compared 7 different dialogue act recognisers in terms of classification accuracy. The comparison was made against gold standard data from a human-labelled corpus. (Semi-Random) is a recogniser choosing a random dialogue act from the Let's Go N-best parsing hypotheses. (Inc_i) is our proposed approach considering a context of i system dialogue acts, and (Ceiling) is a recogniser choosing the correct dialogue act from the Let's Go N-best parsing hypotheses. The latter was used as a gold standard from manual annotations, which reflects the proportion of correct labels in the N-best parsing hypotheses.

We also assessed the impact of ASR N-best information on probabilistic inference. To this end, we compared Bayes nets with a focus on the random variable "*_nbest", which in one case contains induced distributions from data and in the other case contains an equal distribution of slot values. Our hypothesis is that the former setting will lead to better performance.

Figure 3 shows the classification accuracy of our dialogue act recognisers. The first point to notice is that the incorporation of ASR N-best information makes an important difference. The performance of recogniser IncK (K being the number of system dialogue acts) is 66.9% without ASR N-best information and 73.9% with ASR N-best information (the difference is significant[8] at $p < 0.05$). The latter represents a substantial improvement over the semi-random baseline (62.9%) and Lets Go dialogue act recognizer (69%), both significant at $p < 0.05$. A second point to notice is that the differences between Inc_i ($\forall$ i>0) recognisers were not significant. We can say that the use of one system dialogue act as context is as competitive as using a larger set of system dialogue acts. This suggests that dialogue act recognition carried out at early stages (e.g. after the first dialogue act) in an utterance does not degrade recognition performance. The effect is possibly domain-specific and generalisations remain to be investigated.

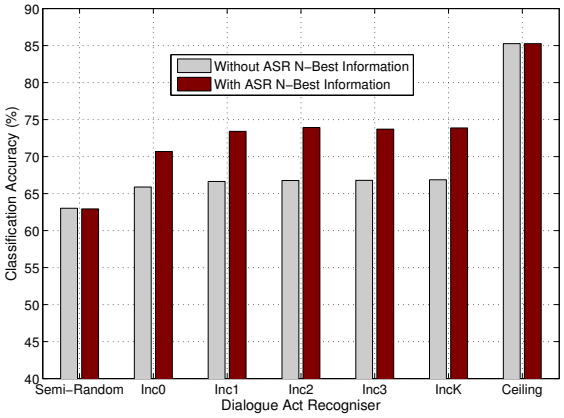

Figure 3: *Bayesian dialogue act recognisers showing the impact of ASR N-best information.*

Generally, we were able to observe that more than half of the errors made by the Bayesian classifiers were due to noise in the environment and caused by the users themselves, which interfered with ASR results. Detecting when users do not convey dialogue acts to the system is therefore still a standing challenge for dialogue act recognition.

5 Conclusion and Future Work

We have described a re-ranking approach for user dialogue act recognition. Multiple Bayesian classifiers are used to rank dialogue acts from a set of dialogue history features and ASR N-best information. Applying our approach to the Let's Go data we found the following: (1) that including ASR N-best information results in improved dialogue act recognition performance; and (2) that competitive results can be obtained from as early as the first system dialogue act, reducing the need to include subsequent ones.

Future work includes: (a) a comparison of our

[6] www.cs.waikato.ac.nz/ml/weka/

[7] www.cs.cmu.edu/~javabayes/Home/

[8] Based on a two-sided Wilcoxon Signed-Rank test.

Bayesian classifiers with other probabilistic models and forms of training (for example by using semi-supervised learning), (b) training dialogue act recognisers in different (multi-modal and multi-task) domains, and (c) dealing with random variables that contain very large domain values.

6 Acknowledgements

This research was funded by the EC FP7 programme under grant agreement no. 287615 (PARLANCE) and no. 270019 (SPACEBOOK).

Sample Re-Ranked User Inputs

User input: "forty six d"

N-Best List of Dialogue Acts	Let's Go Score	Bayesian Score
inform(route=46a)	3.33E-4	1.9236763E-6
inform(route=46b)	1.0E-6	1.5243509E-16
inform(route=46d)	0.096107	**7.030841E-4**
inform(route=46k)	**0.843685**	4.9941495E-10
silence()	NA	0

User input: "um jefferson hills to mckeesport"

N-Best List of Dialogue Acts	Let's Go Score	Bayesian Score
inform(from=mill street)	7.8E-4	3.5998527E-16
inform(from=mission street)	0.015577	3.5998527E-16
inform(from=osceola street)	0.0037	3.5998527E-16
inform(from=robinson township)	0.007292	3.5998527E-16
inform(from=sheraden station)	0.001815	3.1346254E-8
inform(from=brushton)	2.45E-4	3.5998527E-16
inform(from=jefferson)	0.128727	**0.0054255757**
inform(from=mckeesport)	**0.31030**	2.6209198E-4
silence()	NA	0

References

[Bishop2006] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[Bohus and Rudnicky2006] D. Bohus and A. Rudnicky. 2006. A k hypotheses + other" belief updating model. In *AAAI Workshop on Statistical and Empirical Approaches to Spoken Dialogue Systems*.

[Cuayáhuitl and Kruijff-Korbayová2011] H. Cuayáhuitl and I. Kruijff-Korbayová. 2011. Learning human-robot dialogue policies combining speech and visual beliefs. In *IWSDS*, pages 133–140.

[Dethlefs and Cuayáhuitl2011] Nina Dethlefs and Heriberto Cuayáhuitl. 2011. Combining Hierarchical Reinforcement Learning and Bayesian Networks for Natural Language Generation in Situated Dialogue. In *ENLG*, Nancy, France.

[Gambäck et al.2011] Björn Gambäck, Fredrik Olsson, and Oscar Täckström. 2011. Active Learning for Dialogue Act Classification. In *INTERSPEECH*, pages 1329–1332.

[Grau et al.2004] Sergio Grau, Emilio Sanchis, Maria Jose Castro, and David Vilar. 2004. Dialogue Act Classification Using a Bayesian Approach. In *SPECOM*.

[Henderson et al.2012] Matthew Henderson, Milica Gasic, Blaise Thomson, Pirros Tsiakoulis, Kai Yu, and Steve Young. 2012. Discriminative spoken language understanding using word confusion networks. In *SLT*, pages 176–181.

[Keizer and op den Akker2007] Simon Keizer and Rieks op den Akker. 2007. Dialogue Act Recognition Under Uncertainty Using Bayesian Networks. *Natural Language Engineering*, 13(4):287–316.

[Lison2012] Pierre Lison. 2012. Probabilistic dialogue models with prior domain knowledge. In *SIGDIAL Conference*, pages 179–188.

[Mehta et al.2010] Neville Mehta, Rakesh Gupta, Antoine Raux, Deepak Ramachandran, and Stefan Krawczyk. 2010. Probabilistic ontology trees for belief tracking in dialog systems. In *SIGDIAL Conference*, pages 37–46.

[Ng and Jordan2001] Andrew Y. Ng and Michael I. Jordan. 2001. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *NIPS*, pages 841–848.

[Raux et al.2005] Antoine Raux, Brian Langner, Dan Bohus, Alan W. Black, and Maxine Eskenazi. 2005. Let's go public! Taking a Spoken Dialog System to the Real World. In *INTERSPEECH*, pages 885–888.

[Sridhar et al.2009] Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Shrikanth Narayanan. 2009. Combining Lexical, Syntactic and Prosodic Cues for Improved Online Dialog Act Tagging. *Computer Speech & Language*, 23(4):407–422.

[Stolcke et al.2000] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca A. Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialog Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–373.

[Surendran and Levow2006] Dinoj Surendran and Gina-Anne Levow. 2006. Dialog Act Tagging with Support Vector Machines and Hidden Markov Models. In *INTERSPEECH*.

[Tax et al.2000] David M. Tax, Martijn van Breukelen, Robert P. Duin, and Josef Kittler. 2000. Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition*, 33(9):1475–1485, September.

[Thomson et al.2008] Blaise Thomson, Jost Schatzmann, and Steve Young. 2008. Bayesian update of dialogue state for robust dialogue systems. In *ICASSP*, pages 4937–4940.

[Williams2007] Jason D. Williams. 2007. Using particle filters to track dialogue state. In *ASRU*, pages 502–507.

[Wright et al.1999] H. Wright, Massimo Poesio, and Stephen Isard. 1999. Using high level dialogue information for dialogue act recognition using prosodic features. In *Proceedings of an ESCA Tutorial and Research Workshop on Dialogue and Prosody*, pages 139–143, Eindhoven, The Netherlands.

[Zimmermann et al.2005] Matthias Zimmermann, Yang Liu, Elizabeth Shriberg, and Andreas Stolcke. 2005. Toward Joint Segmentation and Classification of Dialog Acts in Multiparty Meetings. In *MLMI*, pages 187–193.