An Evaluation of Alternative Strategies for Implementing Dialogue Policies Using Statistical Classification and Rules

David DeVault, Anton Leuski, and Kenji Sagae
USC Institute for Creative Technologies
12015 Waterfront Drive, Playa Vista, CA 90094
{devault,leuski,sagae}@ict.usc.edu

Abstract

We present and evaluate a set of architectures for conversational dialogue systems, exploring rule-based and statistical classification approaches. In a case study, we show that while a rule-based dialogue policy is capable of high performance if perfect natural language understanding is assumed, a direct classification approach that combines the dialogue policy with NLU has practical advantages.

1 Introduction

In this paper we present and evaluate a set of alternative dialogue system architectures that could be used to implement dialogue policies for conversational characters or virtual humans. The motivation for this work is to improve our understanding of the development costs and performance benefits associated with alternative system architectures for virtual human dialogue systems (Traum et al., 2005; Swartout et al., 2006; Kenny et al., 2009; Jan et al., 2009; Swartout et al., 2010).

We focus on the language processing steps used in a specific virtual human system described in Section 2. We analyze the relationship between Natural Language Understanding (NLU), which maps a user's natural language input to system-specific semantic representations, and Dialogue Management (DM), which executes a dialogue policy that dictates what the virtual human will say or do in response to the user's input.

Traditionally, designing a two step NLU+DM pipeline involves defining semantic representations for the dialogue domain and writing rules that constitute the dialogue policy. This modular design has the benefit of making the DM policy easy to express in explicit rules, but carries the development cost of requiring significant linguistic expertise. Additionally, as we illustrate in this paper, its performance can depend critically on the reliability of the NLU module.

As an alternative, we contrast this design with a direct classification approach that relies only on textual examples and effectively combines the dialogue policy with NLU. In our case study evaluation, we find that this approach offers superior performance, owing to the high frequency of NLU errors in the two step pipeline.

The research presented in this paper extends our previous work. As we summarize in Section 2, this paper relies on the same data set and evaluation metric as DeVault et al. (2011), which reports results for learned policies based on maximum entropy models. In this paper, we add a comparison to a hand-authored policy (Rules) and a new policy based on relevance models (RM). These new policies are described in Section 3. We conclude with some discussion of our new findings.

2 Research Setting and Data Set

We begin by summarizing our research setting, data set, and evaluation metric. We refer the reader to DeVault et al. (2011) for additional details.

We use an existing virtual human scenario designed for Tactical Questioning (TACQ) (Traum et al., 2008), where military personnel interview individuals for information of military value. TACQ characters are designed to be non-cooperative at times. They may answer some of the interviewer's questions, but either lie or refuse to answer others until certain conditions are met (Gandhe et al., 2009). The dialogue policy for a TACQ character is relatively simple in that the character is willing to answer most questions, but correctly implementing the policy requires that certain questions only be answered under certain conditions.

Our work builds on an existing TACQ scenario involving a virtual human called Amani (Gandhe et al., 2009). The user plays the role of a commander whose unit has been attacked by a sniper. The

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1341–1345,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

Lieutenant: *(User Utterance)*
can you tell me what you know of the incident?
NLU Speech Act: `elicit-whq-tellmemoreabouttheincident`
Paraphrases:
- what information do you have about the incident?
- could you please tell me what you saw?
- what can you tell me about the incident?
- can you tell me about the incident?
- please, tell me what you know about the incident
- tell me what you saw, please

Amani: *(System Response, as English text)*
- i saw the shooting. what do you want to know about it?
Other appropriate speech acts, as English text:
- i remember that the gun fire was coming from the window on the second floor of assad's shop.
- what is it you want to know about the incident?

Figure 1: A dialogue turn from the Amani dataset.

user interviews Amani, who was a witness to the incident and has information about the identity of the sniper. Amani is willing to tell the interviewer what she knows, but she will only reveal certain information in exchange for specific promises of safety, secrecy, and monetary compensation (Artstein et al., 2009). Figure 1 provides an excerpt of a user interaction with Amani.

Gandhe et al.'s TACQ system uses speech acts (SAs) to represent the meaning of user and system utterances. In this paper, user utterances are modeled using 46 distinct SA labels. For example, the label `elicit-whq-tellmemoreabouttheincident` is assigned to the user's utterance of *can you tell me what you know of the incident?* in Figure 1. The system also defines a different set of 96 unique SAs (responses) for the Amani character.

We perform our experiments and evaluation using an existing set of 19 annotated Amani dialogues (DeVault et al., 2011). The dialogues were collected through teletype-based role play. Each dialogue turn includes a single user utterance followed by the response chosen by a human role player in the role of Amani. There are a total of 296 turns, for an average of 15.6 turns/dialogue.

The task of Amani's dialogue manager (DM) is to select the most appropriate system SA to use in response to a user utterance. In the experiments reported here, the user's utterance may be provided to the DM either directly as text or using a SA label. We call the DM's decision process a *dialogue policy*. The system builders' intended policy for Amani is detailed in DeVault et al. (2011).

Because Amani has only a fixed set of system responses, the policy problem looks like a traditional classification task. However, there are two sources of uncertainty that complicate the task. Firstly, the mapping between the user's utterance and an appropriate system SA is often one-to-many. In our data set, 6 referees independently linked each user utterance to the best system SA response. In Figure 1, we provide an example in which three different system SAs were selected by the 6 referees. In other cases, up to 6 different system SAs were selected (DeVault et al., 2011). Our first experimental question is therefore: how well can a dialogue policy select an appropriate system SA, if it is provided with an accurate user SA? Would a statistical classification-based policy perform as well as a rule-based policy?

Secondly, the user SAs in the Amani dataset were assigned to the user's utterance by a computational linguist, and we may assume that these "gold" SAs accurately represent the user's intended meaning. In a run-time system, however, the SAs are identified by an automatic NLU module that is likely to introduce errors. It is not obvious *a priori* to what extent the dialogue policy will suffer due to these NLU errors, and our second experimental question is therefore: how well can a policy select an appropriate system SA, if provided with the NLU's hypothesized user SA?

In training the NLU module, as well as our dialogue policies, we can make use of an additional resource in the Amani data set, which is the availability of approximately 6 textual paraphrases for each utterance; see Figure 1 for an example.

As a final empirical question, we consider combining the NLU and DM in a design that classifies user utterances, together with shallow features of the dialogue history, directly into system responses. This approach is similar to the NLU module, but tries to determine system SAs instead of user SAs. This makes unnecessary the labor and knowledge intensive steps of developing the user SA set and annotating utterances with these SAs.

2.1 Evaluation Metric

We evaluate the dialogue policies learned in each of our experimental conditions through 19-fold cross-validation of our set of 19 dialogues. In each fold, we hold out one dialogue and use the remaining 18 dialogues as training data.

To measure the performance of the dialogue policy, we follow the approach of DeVault et al. (2011), which counts an automatically produced system SA as correct if that SA was chosen by at least one referee for that dialogue turn in the data set. We then count the proportion of the correct SAs among all the SAs produced across all 19 dialogues, and use this measure of *weak accuracy* to score competing dialogue policies.

We can use the weak accuracy of one referee, measured against all the others, to establish a performance ceiling for this metric. (We do not expect that an automatic system would outperform a human referee.) This score is .79; see DeVault et al. (2011) for discussion.

3 Experimental Setup

Our experimental setup evaluates three different dialogue policy models: a rule-based approach (Rules), discussed in Section 3.2; a statistical classification technique that uses maximum-entropy classification (MaxEnt), discussed in Section 3.3; and another statistical technique called relevance models (RM), discussed in Section 3.4.

For the Rules approach, the user's utterance is represented in SA form, and we evaluate the performance of the rules using both hand-annotated or "gold" SA (G-SA) as well as automatically assigned NLU SAs (NLU-SA), as described in Section 3.1. For the two statistical policy techniques, MaxEnt and RM, the user's utterance may be represented in SA form or in plain text form. In the latter case, the NLU and DM modules are effectively consolidated into a single classification step.

3.1 NLU

Our NLU module treats the problem of mapping an utterance text to a single SA label as a multi-class classification problem, which it solves using a maximum-entropy model (Berger et al., 1996). The utterance is represented using shallow features such as unigrams and the length of the user utterance (Sagae et al., 2009). Paraphrases of user utterances are included in the training set.

3.2 Rule-based Policy

We developed our rule-based policy (Rules) by manually writing the simple rules needed to implement Amani's dialogue policy. Given a user SA label A_t for turn t, the rules for determining Amani's response R_t take one of three forms:

if $A_t = \mathrm{SA}_i$ then $R_t = \mathrm{SA}_j$
if $A_t = \mathrm{SA}_i \wedge \ \exists k \ A_{t-k} = \mathrm{SA}_l$ then $R_t = \mathrm{SA}_j$
if $A_t = \mathrm{SA}_i \wedge \neg\exists k \ A_{t-k} = \mathrm{SA}_l$ then $R_t = \mathrm{SA}_j$

The first rule form specifies that a given user SA should always lead to a given system response. The second and third rule forms enable the system's response to depend on the user having previously performed (or not performed) a specific SA. For example, Amani will only tell the name of the shooter if the user has previously promised to protect her from danger. If such a promise has not yet been made, she will ask the user to protect her in exchange for the information.

Amani's set of 51 rules was developed in 115 minutes by a computational linguist and system developer. Given the existing set of SAs, the rules were very straightforward to develop.

3.3 MaxEnt Policy

Like the NLU, the MaxEnt policy is based on a multi-class maximum-entropy classifier. It uses text-based features including unigrams and the length of the current and previous user utterance, as well as the SA label for Amani's previous utterance. For experiments in which user utterances are represented as text, the MaxEnt policy is trained using all available paraphrases of user utterances. In experiments in which the user utterance is represented using SA labels rather than text, the paraphrase data is ignored, and the MaxEnt policy is trained using the user SA label in place of the text-based features. In all cases, the MaxEnt policy is trained using all the alternative acceptable Amani SA responses as examples of correct output.

3.4 Relevance Model Policy

The text classification task of assigning the system SAs using either the user SAs or the user text input can be viewed as a cross-language information retrieval (IR) task: we have a fixed collection of system SAs ("documents") and a user's input ("query"), and we need to find the best SA that matches the user's input. This is similar to the task of searching Chinese documents using an English query, where the training data that maps user inputs to the system SAs can be viewed as a "parallel corpus" (Lavrenko et al., 2002).

For our third approach we use the Relevance Model (RM) information retrieval technique first suggested by Lavrenko et al. (2002) and recently adapted to a question-answering task by Leuski et al. (2006). We have experimented with different feature sets and we found that (1) when the text data is not available, the combination of the current user SA and the last system SA is the most

	Policy models		
Utterance Features	Rules	MaxEnt	RM
G-SA	.79	.71	.73
NLU-SA	.58	.57	.60
NLU-SA+Text	-	-	.65
Text	-	.66	.71

Table 1: Weak accuracy results for alternative system architectures.

effective; (2) when both the utterance text and SA are available, the combination of the current user SA and unigram text features from all available paraphrases works the best; and (3) when only the text is available, the unigram word features work well by themselves. We should note that we found it is significantly better to train the model on "gold" SAs even when testing on NLU-SAs. We also observed that integrating the unigram features with the history information in the form of SAs or words from previous utterances tended to over-fit the model, resulting in degraded performance.

4 Results and Discussion

We present our results in Table 1. The highest performance is achieved when "gold" SAs (G-SA) are provided to Rules. Indeed, the weak accuracy of .79 is approximately at the ceiling level of performance observed when one human referee is scored against 5 other human referees. This suggests that, with human-level NLU performance, a hand-authored rule-based policy can effectively implement Amani's intended dialogue policy. However, the table also shows that when automatically-assigned NLU speech acts (NLU-SA) are provided as input to Rules, the performance drops significantly to .58. Note that Rules cannot interpret text representations of user utterances; SA labels are needed, which is a cost of using Rules.

For the MaxEnt policy, a score of .71 is achieved with "gold" SAs, and a lower .57 with run-time SAs. Note that .71 is an inferior performance to the .79 achieved with G-SA/Rules, indicating that MaxEnt does not learn a policy as effective as the hand-authored Rules, even if it is trained and evaluated on gold SA labels. As previously reported in DeVault et al. (2011), a performance of .66 is achieved with the MaxEnt policy when trained on text-based features. It is interesting to see here, however, that this .66 performance is significantly higher than the .58 that is achieved using Rules together with run-time SAs. In fact, the accuracy of the NLU-SA labels in this data set, with respect to the gold SAs, is 53%. Thus, while Rules can achieve very good performance with gold SAs, the high frequency of NLU errors causes a significant degradation in policy performance. Interestingly, the alternative Text/MaxEnt design that combines NLU and DM into a single step ends up performing significantly better (.66).

The RM performance shows a pattern broadly similar to MaxEnt; performance is highest (.73) with gold SAs, and when trained to classify directly from Text to system SAs, performance is significantly better (.71) than NLU-SA/Rules (.58). For RM, we additionally explored using both Text features as well as the NLU-SA label as input features, but observed performance degraded to .65 (presumably due to NLU errors). Our best overall performance not requiring gold SAs, .71, was achieved by Text/RM. Our intuition is that a couple of factors helped RM to outperform the MaxEnt approach: (1) MaxEnt treats word features as binary, while RM explicitly takes into account the word occurrence frequencies; (2) RM is better designed to handle multi-label classification, where a single input instance can be assigned to multiple classes.

5 Conclusion and Future Work

We have presented and evaluated a set of alternative dialogue system architectures in a case study domain. In this domain, we have shown that the theoretical performance that is achievable with a rule-based dialogue policy is high, but that two classification approaches that omit a separate NLU step and directly select system responses perform significantly better. In future work, we plan to address some of the remaining questions, including how these learned policies would perform in live dialogues, how these results would change if NLU performance could be improved, and to what extent this pattern of results would transfer to other domains with more complex NLU and policy requirements.

Acknowledgments

We thank our anonymous reviewers. The project or effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Ron Artstein, Sudeep Gandhe, Michael Rushforth, and David R. Traum. 2009. Viability of a simple dialogue act scheme for a tactical questioning dialogue system. In *DiaHolmia 2009: Proceedings of the 13th Workshop on the Semantics and Pragmatics of Dialogue*, page 43–50, Stockholm, Sweden, June.

Adam L. Berger, Stephen D. Della Pietra, and Vincent J. D. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

David DeVault, Anton Leuski, and Kenji Sagae. 2011. Toward learning and evaluation of dialogue policies with text examples. In *12th annual SIGdial Meeting on Discourse and Dialogue*.

Sudeep Gandhe, Nicolle Whitman, David R. Traum, and Ron Artstein. 2009. An integrated authoring tool for tactical questioning dialogue systems. In *6th Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Pasadena, California, July.

Dusan Jan, Antonio Roque, Anton Leuski, Jackie Morie, and David R. Traum. 2009. A virtual tour guide for virtual worlds. In Zsófia Ruttkay, Michael Kipp, Anton Nijholt, and Hannes Högni Vilhjálmsson, editors, *IVA*, volume 5773 of *Lecture Notes in Computer Science*, pages 372–378. Springer.

Patrick G. Kenny, Thomas D. Parsons, and Albert A. Rizzo. 2009. Human computer interaction in virtual standardized patient systems. In *Proceedings of the 13th International Conference on Human-Computer Interaction. Part IV*, pages 514–523, Berlin, Heidelberg. Springer-Verlag.

Victor Lavrenko, Martin Choquette, and W. Bruce Croft. 2002. Cross-lingual relevance models. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 175–182, Tampere, Finland.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2006. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, Sydney, Australia, July.

Kenji Sagae, Gwen Christian, David DeVault, and David R. Traum. 2009. Towards natural language understanding of partial speech recognition results in dialogue systems. In Short Paper Proceedings of NAACL HLT.

W. Swartout, J. Gratch, R. W. Hill, E. Hovy, S. Marsella, J. Rickel, and D. Traum. 2006. Toward virtual humans. *AI Mag.*, 27(2):96–108.

William R. Swartout, David R. Traum, Ron Artstein, Dan Noren, Paul E. Debevec, Kerry Bronnenkant, Josh Williams, Anton Leuski, Shrikanth Narayanan, and Diane Piepol. 2010. Ada and grace: Toward realistic and engaging virtual museum guides. In Jan M. Allbeck, Norman I. Badler, Timothy W. Bickmore, Catherine Pelachaud, and Alla Safonova, editors, *IVA*, volume 6356 of *Lecture Notes in Computer Science*, pages 286–300. Springer.

David Traum, William Swartout, Jonathan Gratch, Stacy Marsella, Patrick Kenney, Eduard Hovy, Shri Narayanan, Ed Fast, Bilyana Martinovski, Rahul Bhagat, Susan Robinson, Andrew Marshall, Dagen Wang, Sudeep Gandhe, and Anton Leuski. 2005. Dealing with doctors: Virtual humans for non-team interaction training. In *Proceedings of ACL/ISCA 6th SIGdial Workshop on Discourse and Dialogue*, Lisbon, Portugal, September.

David R. Traum, Anton Leuski, Antonio Roque, Sudeep Gandhe, David DeVault, Jillian Gerten, Susan Robinson, and Bilyana Martinovski. 2008. Natural language dialogue architectures for tactical questioning characters. In *Army Science Conference*, Florida, 12/2008.