A Study in How NLU Performance Can Affect the Choice of Dialogue System Architecture

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

Abstract

This paper presents an analysis of how the level of performance achievable by an NLU module can affect the optimal modular design of a dialogue system. We present an evaluation that shows how NLU accuracy levels impact the overall performance of a system that includes an NLU module and a rule-based dialogue policy. We contrast these performance levels with the performance of a direct classification design that omits a separate NLU module. We conclude with a discussion of the potential for a hybrid architecture incorporating the strengths of both approaches.

1 Introduction

Recently computer-driven conversational characters or virtual humans have started finding real-life applications ranging from education to health services and museums (Traum et al., 2005; Swartout et al., 2006; Kenny et al., 2009; Jan et al., 2009; Swartout et al., 2010). As proliferation of these systems increases, there is a growing demand for the design and construction of virtual humans to be made more efficient and accessible to people without extensive linguistics and computer science backgrounds, such as writers, designers, and educators. We are specifically interested in making the language processing and dialogue management components in a virtual human easier for such potential authors to develop. Some system building steps that can be challenging for such authors include annotating the meaning of user and system utterances in a semantic formalism, developing a formal representation of information state, and writing detailed rules that govern dialogue management.

We are generally interested in the extent to which these various authoring steps are necessary in order to achieve specific levels of system performance. In this paper, we present a case study analysis of the performance of two alternative architectures for a specific virtual human. The two architectures, which have been developed and evaluated in prior work (DeVault et al., 2011b; DeVault et al., 2011a), differ substantially in their semantic annotation and policy authoring requirements. We describe these architectures and our evaluation corpus in Section 2. We focus our new analysis specifically on how the overall performance of one of the architectures, which uses a natural language understanding (NLU) module and hand-authored rules for the dialogue policy, depends on the performance of the NLU module. In Section 3, we describe our finding that, depending on the attainable level of NLU accuracy, this modular approach may or may not perform better than a simpler direct classification design that omits a separate NLU module and has a lower annotation and rule authoring burden. In Section 4, we present an initial exploration of whether a hybrid architecture may be able to combine these approaches' strengths.

2 Summary of Data Set and Prior Results

This work is part of an ongoing research effort into techniques for developing high quality dialogue policies using a relatively small number of sample dialogues and low annotation requirements (DeVault et al., 2011b; DeVault et al., 2011a). This section briefly summarizes our prior work and data set.

2.1 Data Set

For our experiments we use the dataset described in (DeVault et al., 2011b). It contains 19 Wizard of Oz dialogues with a virtual human called Amani (Gandhe et al., 2009). The user plays the role of an Army commander whose unit has been attacked by a sniper. The user interviews Amani, who was a witness to the incident and has some information about the sniper. Amani is willing to tell the interviewer what she knows, but she will only reveal certain information in exchange for promises of safety, secrecy, and money (Artstein et al., 2009).

Each dialogue turn in the data set includes a single user utterance followed by the response chosen by a human Amani role player. There are a total of 296 turns, for an average of 15.6 turns/dialogue. User utterances are modeled using 46 distinct speech act (SA) labels. The dataset also defines a different set of 96 unique SAs (responses) for Amani. Six external referees analyzed each user utterance and selected a single character response out of the 96 SAs. Thus the dataset defines a one-to-many mapping between user utterances and alternative system SAs.

2.2 Evaluation Metric

We evaluate the dialogue policies in our experiments through 19-fold cross-validation of our 19 dialogues. In each fold, we hold out one dialogue and use the remaining 18 as training data. To measure policy performance, we count an automatically produced system SA as correct if that SA was chosen by the original wizard or at least one external referee for that dialogue turn. 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 dialogue policies.

We can use the weak accuracy of one referee, measured against all the others, to establish a performance ceiling for this metric. This score is .79; see DeVault et al. (2011b).

2.3 Baseline Systems

We consider two existing baseline systems in our experiments here. The first system (Rules-NLU-SA) consists of a statistical NLU module that maps a user utterance to a single user SA label, and a rule-based dialogue policy hand-crafted by one of the authors. The NLU uses a maximum-entropy model (Berger et al., 1996) to classify utterances as one of the user SAs using shallow text features. Training this model requires a corpus of user utterances that have been semantically annotated with the appropriate SA.

We developed our rule-based policy 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:

$(a)\text{if}\, A_t = \text{SA}_i \,\text{then}\, R_t = \text{SA}_j$
$(b)\text{if}\, A_t = \text{SA}_i \wedge \quad \exists k A_{t-k} = \text{SA}_l \,\text{then}\, R_t = \text{SA}_j$
$(c)\text{if}\, A_t = \text{SA}_i \wedge \neg\exists k A_{t-k} = \text{SA}_l \,\text{then}\, R_t = \text{SA}_j$

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. One the system developers, who is also a computational linguist, created the current set of 42 rules in about 2 hours. There are 30 rules of form (a), 6 rules of form (b), and 6 rules of form (c).

The second baseline system (RM-Text) is a statistical classifier that selects system SAs by analyzing shallow features of the user utterances and system responses. We use the Relevance Model (RM) approach pioneered by Lavrenko et al. (2002) for cross-lingual information retrieval and adapted to question-answering by Leuski et al. (2006). This method does not require semantic annotation or rule authoring; instead, the necessary training data is defined by linking user utterances directly to the appropriate system responses (Leuski and Traum, 2010).

Table 1 summarizes the performance for the baseline systems (DeVault et al., 2011a). The NLU module accuracy is approximately 53%, and the weak accuracy of .58 for the corresponding system (Rules-NLU-SA) is relatively low when compared to the RM system at .71. For comparison we provide a third data point: for Rules-G-SA, we assume that our NLU is 100% accurate and always returns the correct ("gold") SA label. We then run the rule-based dialogue policy on those labels. The third column (Rules-G-SA) shows the resulting weak accuracy value, .79, which is comparable to the weak accuracy score achieved by the human referees (DeVault et al., 2011b).

Rules-NLU-SA	RM-Text	Rules-G-SA
.58	.71	.79

Table 1: Weak accuracy results for baseline systems.

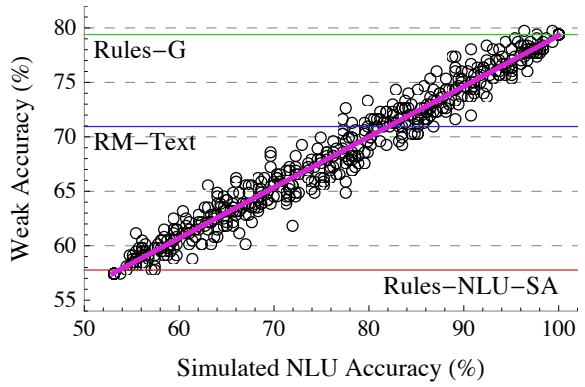

Figure 1: Weak accuracy of the Rules system as a function of simulated NLU accuracy.

3 NLU Accuracy and System Performance

We conducted two experiments. In the first, we studied the effect of NLU accuracy on the performance of the Rules-NLU-SA system. One of our goals was to find how accurate the NLU would have to be for the Rules-NLU-SA system to outperform RM-Text.

To investigate this, we simulated NLU performance at different accuracy levels by repeatedly sampling to create a mixture of the SAs from the trained NLU classifier and from the correct (gold) set of SAs. Specifically, we set a fixed value p ranging from 0 to 1 and then iterate over all dialogue turns in the held out dialogue, selecting the the correct SA label with probability p or the trained NLU module's output with probability $1 - p$. Using the sampled set of SA labels, we compute the resulting simulated NLU accuracy, run the Rules dialogue policy, and record the weak accuracy result. We repeat the process 25 times for each value of p. We let p range from 0 to 1 in increments of .05 to explore a range of simulated accuracy levels.

Figure 1 shows simulated NLU accuracy and the corresponding dialogue policy weak accuracy as a point in two dimensions. The points form a cloud with a clear linear trend that starts at approximately 53% NLU accuracy where it intersects with the Rules-NLU-SA system performance and then goes up to the Rules-G performance at 100% NLU accuracy. The correlation is strong with $R^2 = 0.97$.[1]

The existence of a mostly linear relationship comports with the fact that most of the policy rules (30 of 42), as described in Section 2.3, are of form (a). For such rules, each individual correct NLU speech act translates directly into a single correct system response, with no dependence on the system having understood previous user utterances correctly. In contrast, selecting system responses that comply with rules in forms (b) and (c) generally requires correct understanding of multiple user utterances. Such rules create a nonlinear relationship between policy performance and NLU accuracy, but these rules are relatively few in number for Amani.

The estimated linear trend line (in purple) crosses the RM-Text system performance at approximately 82% NLU accuracy. This result suggests that our NLU component would need to improve from its current accuracy of 53% to approximately 82% accuracy for the Rules-NLU-SA system to outperform the RM-Text classifier. This represents a very substantial increase in NLU accuracy that, in practice, could be expected to require a significant effort involving utterance data collection, semantic annotation, and optimization of machine learning for NLU.

4 Hybrid System

In our second experiment we investigated the potential to integrate the Rules-NLU-SA and RM-Text systems together for better performance. Our approach draws on a confidence score θ from the NLU maximum-entropy classifier; specifically, θ is the probability assigned to the most probable user SA.

Figure 2 shows an analysis of NLU accuracy, Rules-NLU-SA, and RM-Text that is restricted to those subsets of utterances for which NLU confidence θ is greater than or equal to some threshold τ. Two important aspects of this figure are (1) that raising the minimum confidence threshold also raises the NLU accuracy on the selected subset of utterances; and (2) that there is a threshold NLU confidence level beyond which Rules-NLU-SA seems to

[1]This type of analysis of dialogue system performance in terms of internal component metrics is somewhat similar to the regression analysis in the PARADISE framework (Walker et al., 2000). However, here we are not concerned with user satisfaction, but are instead focused solely on the modular system's ability to reproduce a specific well-defined dialogue policy.

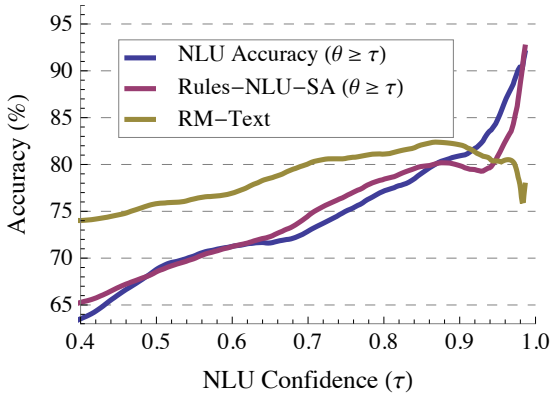

Figure 2: Weak accuracy of Rules-NLU-SA and RM-Text on utterance subsets for which NLU confidence $\theta \geq \tau$. We also indicate the corresponding NLU accuracy at each threshold. In all cases a rolling average of 30 data points is shown to more clearly indicate the trends.

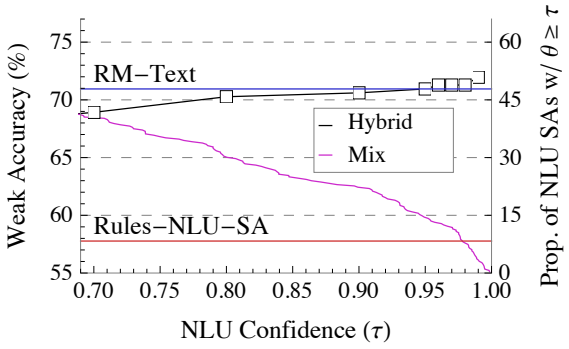

Figure 3: Weak accuracy of the Hybrid system as a function of the NLU confidence score.

outperform RM-Text. This confidence level is approximately 0.95, and it identifies a subset of user utterances for which NLU accuracy is 83.3%. These results therefore suggest that NLU confidence can be useful in identifying utterances for which NLU speech acts are more likely to be accurate and Rules-NLU-SA is more likely to perform well.

To explore this further, we implemented a hybrid system that chooses between Rules-NLU-SA or RM-Text as follows. If the confidence score is high enough ($\theta \geq \tau$, for some fixed threshold τ), the Hybrid system uses the NLU output to run the Rules dialogue policy to select the system SA; otherwise, it discards the NLU SA, and applies the RM classifier to select the system response directly.

Figure 3 shows the plot of the Hybrid system performance as a function of the threshold value τ. We see that with sufficiently high threshold value ($\tau \geq 0.95$) the Hybrid system outperforms both the Rules-NLU-SA and the RM-Text systems. The second line, labeled "Mix" and plotted against the secondary (right) axis, shows the proportion of the NLU SAs with the confidence score that exceed the threshold ($\theta \geq \tau$). It indicates how often the Hybrid system prefers the Rules-NLU-SA output over the RM-Text system output. We observe that approximately 42 of the NLU outputs over all 296 dialogue turns (15%) have confidence values $\theta \geq 0.95$. However, for most of these dialogue turns the outputs for the Rules-NLU-SA and RM-Text dialogue policies are the same. While we observe a small improvement in the Hybrid system weak accuracy values over the RM-Text system at thresholds of 0.95 and higher, the difference is not statistically significant.

Despite the lack of statistical significance in the initial Hybrid results in this small data set, we interpret the complementary evidence from both experiments, which support the potential for Rules-NLU-SA to perform well when NLU accuracy is high, and the potential for a hybrid system to identify a subset of utterances that are likely to be understood accurately at run-time, as indicating that a hybrid design is a promising avenue for future work.

5 Conclusions and Future Work

We presented a case study analysis of how the level of performance that is achievable in an NLU module can provide perspective on the design choices for a modular dialogue system. We found that NLU accuracy must be substantially higher than it currently is in order for the Rules-NLU-SA design, which carries a greater annotation and rule authoring burden, to deliver better performance than the simpler RM-Text design. We also presented evidence that a hybrid architecture could be a promising direction.

Acknowledgments

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. 2011a. An evaluation of alternative strategies for implementing dialogue policies using statistical classification and rules. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 1341–1345, Nov.

David DeVault, Anton Leuski, and Kenji Sagae. 2011b. Toward learning and evaluation of dialogue policies with text examples. In *Proceedings of the 12th annual SIGdial Meeting on Discourse and Dialogue*, pages 39–48.

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 and David Traum. 2010. NPCEditor: A tool for building question-answering characters. In *Proceedings of The Seventh International Conference on Language Resources and Evaluation (LREC)*.

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.

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.

Marilyn Walker, Candace Kamm, and Diane Litman. 2000. Towards developing general models of usability with PARADISE. *Nat. Lang. Eng.*, 6(3-4):363–377.