Toward Learning and Evaluation of Dialogue Policies with Text Examples

David DeVault and **Anton Leuski** and **Kenji Sagae**
Institute for Creative Technologies
University of Southern California
Playa Vista, CA 90094
{devault,leuski,sagae}@ict.usc.edu

Abstract

We present a dialogue collection and enrichment framework that is designed to explore the learning and evaluation of dialogue policies for simple conversational characters using textual training data. To facilitate learning and evaluation, our framework enriches a collection of role-play dialogues with additional training data, including paraphrases of user utterances, and multiple independent judgments by external referees about the best policy response for the character at each point. As a case study, we use this framework to train a policy for a limited domain tactical questioning character, reaching promising performance. We also introduce an automatic policy evaluation metric that recognizes the validity of multiple conversational responses at each point in a dialogue. We use this metric to explore the variability in human opinion about optimal policy decisions, and to automatically evaluate several learned policies in our example domain.

1 Introduction

There is a large class of potential users of dialogue systems technology who lack the background for many of the formal modeling tasks that typically are required in the construction of a dialogue system. The problematic steps include annotating the meaning of user utterances in some semantic formalism, developing a formal representation of information state, writing detailed rules that govern dialogue management, and annotating the meaning of system utterances in support of language generation, among other tasks.

In this paper, we explore data collection and machine learning techniques that enable the implementation of domain-specific conversational dialogue policies through a relatively small data collection effort, and without any formal modeling. We present a case study, which serves to illustrate some of the possibilities in our framework. In contrast to recent work on data-driven dialogue policy learning that learns dialogue behavior from existing data sources (Gandhe and Traum, 2007; Jafarpour et al., 2009; Ritter et al., 2010), we address the task of authoring a dialogue policy from scratch with a specific purpose, task and scenario in mind. We examine the data collection, learning and evaluation steps.

The contributions of this work include a data collection and enrichment framework without formal modeling, and the creation of dialogue policies from the collected data. We also propose a framework for evaluating learned policies. We show, for the scenario in our case study, that these techniques deliver promising levels of performance, and point to possible future developments in data-driven dialogue policy creation and evaluation.

2 Case study

For our case study we selected an existing dialogue system scenario designed for Tactical Questioning training (Traum et al., 2008). The character targeted in our study, Amani, is modeled closely after the Amani Tactical Questioning character described by Gandhe et al. (2009) and Artstein et al. (2009). Tactical Questioning dialogues are those in which small unit military personnel, usually on patrol, hold conversations with individuals to produce information of military value. A tactical questioning dialogue

Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 39–48,
Portland, Oregon, June 17-18, 2011.

system is a simulation training environment where virtual characters play the role of a person being questioned. Tactical questioning characters are designed to be non-cooperative at times. They may answer some of the interviewers questions in a cooperative manner, but may refuse to answer other questions, or intentionally provide incorrect answers. Therefore the interviewer is encouraged to conduct the interview in a manner that induces cooperation from the character: building rapport with the character, addressing their concerns, making promises and offers, as well as threatening or intimidating the character; the purpose of the dialogue system is to allow trainees to practice these strategies in a realistic setting (Gandhe et al., 2009).

This type of scenario is a good testbed for our proposed learning and evaluation framework, since it involves both flexible conversational choices and well-defined constraints regarding the disclosure of specific information. In the Amani scenario, the user plays the role of a commander of a small military unit in Iraq whose unit had been attacked by sniper fire. The user interviews a character named Amani who was a witness to the incident and is thought to have some information about the identity of the attackers. Amani is willing to tell the interviewer everything she knows provided that the user promises her safety, secrecy, and small monetary compensation for the information (Artstein et al., 2009).

An exhaustive formal definition of Amani's ideal dialogue policy might include a large number of rules covering a wide range of user utterance types. The key constraints for the training simulation, however, can be stated simply with a few rules governing the release of five pieces of information that Amani knows. Amani will only reveal one of these pieces of information if a precondition is met. Table 1 shows how certain information relates to each of the preconditions in Amani's dialogue policy. Amani can only reveal a fact from the first column if the user promised her an item from the second column. For example, Amani can only tell the user the shooter's name if the user promised her safety. If the user has not promised safety, Amani will ask him for safety. If the user refuses to promise safety, Amani will either decline to answer the question or lie to the interviewer. Amani does keep track of the user's promises and once she is promised safety, she would not ask for it again.

information	precondition
about shooter's name	safety
about shooter's description	safety
about shooter's location	secrecy
about the occupant of the shop	secrecy
about shooter's daily routine	money

Table 1: Amani's dialogue policy.

While the key constraints for Amani's policy, as summarized in Table 1, may be easily expressed in terms of rules involving dialogue-acts, the rest of Amani's behavior is more open-ended and underspecified. Ideally, the system designers would like for the character to obey conversational conventions (such as responding appropriately to greetings, thankings, etc.). Her responses to other user utterances should match human intuition about what a good response would be, but specific responses are not generally dictated by the goals for the training simulation. There is therefore room for some flexibility, and also for the character to reply that she does not understand. Of course, her conversational repertoire is inevitably limited by the available authoring and development effort as well as language processing challenges.

3 Data collection

The exponential number of possible utterances and dialogue paths in even a simple conversational dialogue scenario such as the Amani scenario suggests that learning acceptable dialogue behavior from surface text examples without annotation or formal modeling would require a seemingly insurmountable quantity of dialogues to serve as training data. We address this problem in a data collection framework with four main characteristics: (1) we sidestep the problem of learning natural language generation by using a fixed predefined set of utterances for the Amani character. This so-called "utterance selection" approach has been used in a number of dialogue systems (Zukerman and Marom, 2006; Sellberg and Jnsson, 2008; Kenny et al., 2007, for example) and often serves as a reasonable approximation to generation (Gandhe and Traum, 2010); (2) we collect dialogues from human participants who play the parts of Amani and the commander in a *structured role play* framework (Section 3.1); (3) we enrich the dialogues collected in the structured role play step with additional paraphrases for the utterances of the commander, in an attempt to deal with large variability of natural language input, even for a limited domain conversational dialogue scenario (Section 3.2); (4) we further augment the existing dialogue data by adding acceptable alternatives to the dialogue acts of the Amani role through the use of *external referees* (Section 3.3).

Our data collection procedure is designed to capture the necessary information for learning dialogue policies and evaluating their quality by approximating the exponentially large dialogue variability while keeping the data collection effort tractable.

3.1 Structured role play

To examine the hypothesis that dialogue policies such as Amani's can be learned from examples without explicit rules or any kind of formal modeling, we collected dialogue data through a constrained form of role play, which we call *structured role play*, where the person playing the role of Amani is encouraged, whenever possible, to only use utterances from a fixed set. Each utterance in the available set of Amani replies corresponds roughly to one of the dialogue acts (consisting of an illocutionary force and some semantic content) described by Artstein et al. (2009) for their version of the Amani character.

The players in the roles of Amani and the commander take turns producing one utterance at a time, each in a separate terminal. The commander player, who receives a natural language description of the scenario and the goal of the commander, enters utterances through a teletype (chat) interface. The Amani player, who receives a natural language description of the scenario and of Amani's dialogue policy, chooses an utterance from a list for each dialogue turn. The Amani player is encouraged to use an utterance from this list whenever possible; however, for user utterances that the Amani player judges cannot possibly be handled by any existing response, a new response can be authored (as English text) and immediately used in the role play. Each player sees the other's utterance as text in their own terminal. This closely resembles a Wizard-of-Oz setup, with they key difference being that both dialogue participants believe they are interacting with another person, which is in fact the case, and the idea of a wizard controlling a system is not part of the exercise. However, because the Amani player is encouraged to limit Amani's responses to a fixed utterance set, and the dialogue is constrained to a strict turn-taking setup that interleaves utterances from each participant, the situation also differs from conventional role play.

We collected a total of 19 dialogues and 296 utterances for Amani, for an average of 15.6 Amani utterances per dialogue.

3.2 Paraphrase generation

The dialogues collected through structured role play are intended for serving as training data from which Amani's dialogue policy can be learned. However, to cover the natural language variability with which dialogue acts from the commander can be expressed would require a much larger number of dialogues than it would be practical to collect, since a learned system that deals only with the surface text in the dialogues would need to deal both with the dialogue policy and natural language understanding for the scenario. Instead, we require only that the dialogues collected cover the desired dialogue acts for the player role in the scenario. To address the language understanding problem (indirectly), we attempt to cover the variability of expression of these dialogue acts through the collection of paraphrases for the commander utterances in the set of dialogues.

For each commander utterance in the 19 dialogues we collected, we had annotators create a set of paraphrases. In creating paraphrases, annotators were asked to take not just the original utterance into account, but also its context in the dialogue. We did not specify a fixed number of paraphrases per utterance, but instead asked for as many paraphrases as the annotator could quickly think of.

Figure 1 exemplifies the paraphrases created during this process, for a target user utterance of *can you tell me what you know of the incident?*. For this utterance, a total of 6 paraphrases were generated.

We used a total of 9 annotators, who created paraphrases for the 296 utterances in the 19 dialogues. Most annotators were responsible for no more than two dialogues, and took on average less than 30 minutes per dialogue. The average number of para-

Previous Dialogue History	
Lieutenant:	*hi amani*
Amani:	*hello.*
Lieutenant:	*how are you doing?*
Amani:	*fine thank you.*
Lieutenant:	*thank you for meeting with me*
Amani:	*you're welcome.*
Target User Utterance	
Lieutenant:	*can you tell me what you know of the incident?* **Paraphrases:** *please tell me what information 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*
System Response	
Amani:	*i saw all the shooting from my window. what do you want to know about it?* **External Referees:** (3 referees) *i saw all the shooting from my window. what do you want to know about it?* (2 referees) *i remember that the gun fire was coming from the window on the second floor of assad's shop. the shop is only one story but there are apartments on top of the shop.* (1 referee) *what is it you want to know about the incident?*

Figure 1: An enriched dialogue turn from an Amani structured role play.

phrases collected per user utterance was 5.5.

Our 9 annotators had differing backgrounds, ranging from transcribers and summer interns to experienced NLP researchers. It should be noted that all had at least some experience working with natural language processing technologies. In future work, we would like to explore using less experienced annotators for paraphrasing.

3.3 External referee annotation

Although the paraphrase generation step helps with coverage of the language used by the commander in our scenario, the combination of the original dialogues collected through structured role play and the paraphrases do not address one crucial issue in learning of data-driven dialogue policies, and their automated evaluation: at each turn, a dialogue participant has multiple valid dialogue acts that can be performed, not a single correct one. In other words, given the same dialogue history up to a given point, multiple human dialogue participants following the same underspecified policy may choose different dialogue acts to continue the dialogue, and each of these different choices may be perfectly acceptable and coherent. This is one of main challenges in creation and evaluation of data-driven policies, since the exponentially many acceptable dialogue paths are both difficult to model explicitly, and difficult to recognize automatically when performed during testing. Of course, the degree to which this is a practical problem in a specific dialogue scenario depends on several factors, including how underspecified the targeted dialogue policy is. In our case study, the policy has a high level of underspecification, since only behaviors related to the information in Table 1 are mentioned directly, and even those are only described in natural language, without formal rigor. The rest of the policy dictates only that human players in the part of Amani act according to their commonsense in playing the role of the Amani character. However, we limit the otherwise potentially infinite possibilities for dialogue behavior by strongly encouraging the Amani player to perform only one of a set of predefined utterances corresponding to certain dialogue acts in the scenario. In our experiments, the number of utterances available for Amani was 96.

We first investigate this issue by attempting to characterize the amount of human variation in the choice of one of the 96 available dialogue acts at any given point in a dialogue. To this end, we introduce the idea of the *external referee*, who essentially provides a "second opinion" for dialogue acts performed by the original role player. The external referee annotation task works as follows: (1) Starting with an existing dialogue containing n utterances

$\langle u_1, u_2, ..., u_n \rangle$ for the participant whose utterances will be externally refereed (one of the dialogues collected through structured role play, in our case study, where we externally referee the Amani utterances), produce n dialogue histories $h_1, h_2, ..., h_n$, with each h_i consisting of every utterance from each dialogue participant from the beginning of the dialogue down to, but not including, the i^{th} utterance in the dialogue. (2) For each dialogue history h_i, the external referee (who must not be the person who played a part in the original dialogue) chooses an utterance u'_i from the choices available for the scenario, without knowledge of the original utterance u_i in the dialogue from which the history was produced.

Figure 1 provides an example of the choices made by 6 external referees for a single target user utterance. Given the previous dialogue history and the target user utterance (*can you tell me what you know of the incident?*), each external referee independently chose a single best utterance for the character to respond with. In the example in the figure, it can be seen that 3 of the 6 external referees chose the same response as the original Amani player, asserting that Amani did indeed witness the incident and asking what the commander would like to know. The other three chose alternative responses; two of these selected a response asserting information about where the gun fire was coming from, while a third referee chose a response simply asking what the commander would like to know. It is important to note that all three of these alternative responses would be acceptable from a design and training perspective.

In this annotation task, the task is not to provide alternative dialogues, but simply one character response to each individual utterance, assuming the fixed history of the original dialogue. In other words, the annotator has no control or impact over the dialogue history at any point, and provides only additional reference utterances for possible immediate continuations for each dialogue history. It is for this reason we call the annotator an external referee.

Annotations from multiple external referees for the dialogues collected through structured role play do not result in a representation of the lattice of the many possible dialogue paths in the scenario, but rather an approximation that represents the possible options in the immediate future of a given dialogue history. The main difference is that the available histories are limited to those in the original dialogues from structured role play. While this may be a limiting factor if one attempts to model dialogue behavior based on entire dialogue histories, since the available histories represent only a very sparse sample of the space of valid histories, it is possible that good approximate models can be achieved with factorization of dialogues by sequences of a fixed number of consecutive turns, e.g. a model that makes a second-order Markov assumption, considering only the previous two turns in the dialogue as an approximation of the entire history (Gandhe and Traum, 2007). This is in a way the same approximation used in n-gram language models, but at the level of granularity of sentences, rather than words.

We collected annotations from 6 different external referees, with each individual referee annotating the entire set of 19 dialogues, and taking on average about two hours to complete the annotation of the entire set. All of our external referees were very familiar with the design of the Amani character, and most had natural language processing expertise.

4 Evaluation of dialogue policies with multiple external referees

4.1 External referee agreement

The dialogues and external referee annotations collected using the procedure described in Section 3 provide a way to characterize the targeted policy with respect to human variability in choosing utterances from a fixed set, since the annotations include the choices made by multiple external referees.

From the annotations of utterances chosen for Amani in our 19 dialogues, we see that human annotators agree only 49.2% of the time when choosing an utterance in the external referee framework. That is, given the same dialogue history, we expect that two human role players would agree on average slightly less than 50% of the time on what the next utterance should be[1].

Based on this level of pairwise agreement, one might conclude that using these data for either policy learning or policy evaluation is a lost cause. How-

[1] This represents the averaged agreement over all pairs of external referees.

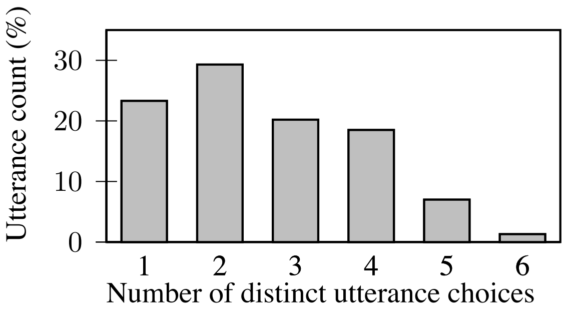

Figure 2: Distribution in number of distinct choices by external referees

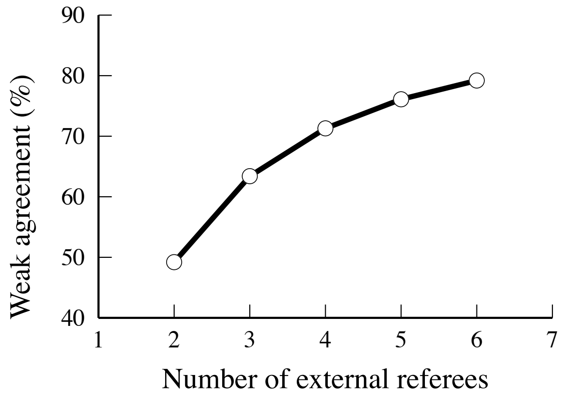

Figure 3: Weak agreement between external referees

ever, this result does not necessarily indicate that human raters disagree on what the correct choice is; it is more likely to reflect that there are in fact multiple "correct" (acceptable) choices, which we can capture through multiple annotators.

The annotations from multiple external referees in our case study support this view: Figure 2 shows the number of distinct utterance choices made by each of the six external referees for each specific utterance in the 19 dialogues collected through structured role play. Each external referee chooses only one utterance (out of 96 options) per Amani turn in the 19 dialogues. Over the 296 Amani utterances in the entire set of dialogues, all six referees agreed unanimously on their utterance choice only 23.3% of the time. The most frequent case, totaling almost 30% of all utterances, was that the set composed by the single choice from each of the six wizards for an utterance had exactly two distinct elements. For only 1.3% of the 296 utterances did that set contain the maximum number of distinct elements (six), indicating complete disagreement among the external referees. We note that, in this case, very low agreement to complete disagreement reflects a situation in dialogue where it is likely that there are many dialogue act choices considered acceptable by the collective body of external referees. In our scenario, there were at most two choices from the six referees for more than 50% of the Amani turns, indicating that in the majority of the cases there is only a small set of acceptable dialogue acts (from the 296 available), while five or more options were chosen for less than 10% of all Amani turns.

For a more direct characterization of dialogue scenarios, and also for the purposes of evaluation, we now define a metric that reflects overall agreement in a group of external referees. Instead of comparing one choice from a single referee to another single choice, we instead check for membership of a single choice c_{ij} from a single referee R_i for utterance u_j in the set of choices $\{c_{kj} | k \neq i\}$ from all of the other referees $\{R_k | k \neq i\}$. In the positive case, we say that R_i *weakly agrees* with the rest of the raters $\{R_k | k \neq i\}$ on the annotation of utterance u_j. We define the *weak agreement* agr_n for a set of N external referees over a set of m utterances to be rate at which each rater R_i weakly agrees with the $n - 1$ raters $\{R_k | k \neq i\}$, for all integer values of i ranging from 1 to N, inclusive. Intuitively, weak agreement reflects two important questions: (1) how often is the choice of a referee supported by the choice of at least one more referee? and (2) given a set of $n - 1$ referees, how much new information (in the form of unseen choices) should I expect to see from a new n^{th} referee? Figure 3 addresses these questions for the scenario in our case study by showing the weak agreement figures obtained for sets of increasing numbers of external referees, from 2 to 6. Each point in the graph corresponds to the average of the weak agreement values obtained for all possible ways of holding out one external referee R_i, and computing the weak agreement between R_i and the other referees, assuming an overall pool containing the given number of external referees.

We note that with the dialogue act choices of a single person, coverage of the possible acceptable options is quite poor, corresponding only to an average of 50% of the choices made by another person.

The coverage increases rapidly as two more external referees are added, and more slowly, although still steadily from there. The rightmost point in Figure 3 indicates that with a set of five external referee we should expect to cover almost 80% of the choices of a sixth referee.

4.2 Dialogue policy evaluation with multiple external referees

The weak agreement metric defined in the previous section can be used to measure the quality of automatically learned policies, and to provide insight into how a learned policy compares to human-level performance. Because it recognizes the validity of multiple responses, the weak agreement metric can help distinguish true policy errors from policy choices that are consistent with the intuitions of at least some human referees about what the character should say.

In particular, given the choices made by five external referees for our 19 Amani dialogues, we can expect their choices to cover about 80% of the choices a sixth person would make for what Amani should say at each turn in these dialogues. (I.e., we know that the weak agreement among a group of six human referees is about 80% for this Amani scenario.)

We proceed to rate the quality of an automatic policy by computing a one-vs-others version of weak agreement—intuitively treating our policy as if it were such a "sixth person", and comparing it to the other five. Instead of computing the average weak agreement for referees randomly selected from an entire group, as in the previous section, to evaluate a policy, we compute its weak agreement compared to the combined set of human external referees, as follows. For every system utterance u_j in our set of role play dialogues, a given automatic policy P is used to select a response c'_j (corresponding to a dialogue act in the domain). We then check for membership of c'_j in the set that contains only and all dialogue act choices c_{kj} for k ranging from 1 to N, inclusive, where N is the number of external referees and c_{kj} corresponds to the k^{th} referee's choice for the j^{th} utterance. Another way to interpret this evaluation metric is to consider it a form of accuracy that computes the number of correct choices made by the policy divided by the total number of choices made by the policy, where a choice is considered "correct" if it matches any of the external referees' choices for a specific utterance. For this reason, we refer to this evaluation-focused one-vs-all version of weak agreement as *weak accuracy*.

Based on the definition above, an automatic policy with quality indistinguishable from that of a person choosing utterances for the Amani character would have a weak accuracy of about 80% or higher when measured using a set of five external referees. We see then that this metric is far from perfect, since it cannot rank two policies with weak accuracy levels of, say, 80% and 90%. It is also possible for a policy that results in dialogue behavior noticeably inferior to that of a human referee to be rated at the same weak accuracy value for a human referee (80%). In practice, however, weak accuracy with five or six external referees has far greater power for discriminating between policies of varying quality, and ranking them correctly, than a naive version of accuracy, which corresponds to weak accuracy using a single referee. Furthermore, the addition of only a few more external referees would very likely increase the efficacy of the weak agreement metric.

Despite the shortcomings of weak accuracy as a metric for evaluation of quality of dialogue policies, it opens up a wide range of opportunities for development of learned policies. Without an automated metric, development of such techniques can be only vaguely incremental, relying on either costly or, more likely, infrequent human evaluations with results that are difficult to optimize toward with current machine learning techniques. The use of imperfect automated metrics in situations where ideal metrics are unavailable or are impractical to deploy is fairly common in natural language processing. PARSEVAL (Abney et al., 1991), commonly used for parser evaluation, and BLEU (Papineni et al., 2002), commonly used in machine translation, are two examples of well-known imperfect metrics that have been the subject of much criticism, but that are widely agreed to have been necessary for much of the progress enjoyed by their respective fields. Unlike BLEU, however, which has been shown to correlate with certain types of human judgment on the quality of machine translation systems, our notion of weak accuracy has not yet been demonstrated to correlate with human judgments on the quality of dialogue policies, and as such it is only hypothesized to have this property. We leave this important step of validation as future work.

5 Learning dialogue policies from examples without formal modeling

Equipped with a dataset with 19 dialogues in the Amani scenario (including paraphrases for the unconstrained commander utterances, and external referee annotations for the constrained Amani utterances), and an automatic evaluation framework for distinguishing quality differences in learned policies, we now describe our experiments on learning dialogue policies from data collected in structured role play sessions, and enriched with paraphrases and external referee annotations.

In each of our experiments we attempt to learn a dialogue policy as a maximum entropy classifier (Berger et al., 1996) that chooses one utterance out of the 96 possible utterances for Amani after each commander utterance, given features extracted from the dialogue history. This policy could be integrated in a dialogue system very easily, since it chooses system utterances directly given previous user and system utterances. We evaluate the dialogue policies learned in each experiment through 19-fold cross-validation of our set of 19 dialogues: in each fold, we hold out one dialogue (and all of its related information, such as external referee annotations and user utterance paraphrases) and use the remaining 18 dialogues as training data.

5.1 Learning from examples

Using only the dialogues collected in structured role play sessions, and no additional information from external referees or paraphrases, we train the maximum entropy classifier to choose a system utterance s_i based on features extracted from the two previous user utterances u_i and u_{i-1} and the previous system utterance s_{i-1}. The features extracted from these utterances are the words present in each user utterance, and the complete text of each system utterance. Low frequency words occurring fewer than 5 times in the corpus are excluded.

The weak accuracy for this simple policy is 43%, a low value that indicates that for more than half its turns the policy chooses an utterance that was not chosen by any of the referees, giving us a reasonable level of confidence that this policy is of poor quality.

5.2 Enhanced training with external referees

The next experiment expands the training set available to the maximum entropy classifier by adding training instances based on the utterances chosen by the external referees. For each of the training instances (target utterance coupled with features from u_i, s_{i-1} and u_{i-1}) we add six new training instances, each using the same features as the original training instance, but replacing the target class with the choice made by an external referee. Note that this creates identical training instances for cases when the same utterance is chosen by multiple annotators, which has the effect of weighting training examples. With the additional information, weak accuracy for this policy improves to 56%, which is a large gain that still results in a mediocre dialogue policy.

5.3 Expanding training examples with paraphrases

To help determine how much of difficulty in our policy learning task is due to the related problem of natural language understanding (NLU), and how much is due to modeling dialogue behavior regardless of NLU, we performed manual annotation of dialogue acts for the user utterances, and trained a policy as in the previous section, but using manually assigned dialogue acts instead of the words for user utterances in the dialogue history. With this gold-standard NLU, weak accuracy improves from 56% to 67%, approaching the level of human performance, and already at a level where two out of every three choices made by the learned policy matches the choice of a human referee.

To bridge the gap between learning purely from surface text (with no formal modeling) and learning from manually assigned dialogue acts specifically designed to capture important information in the scenario, we turn to the paraphrases collected for user utterances in our 19 dialogues. These paraphrases are used to create additional synthetic training material for the classifier, as follows: for each training instance produced from a chosen system utterance s_i and previous utterances u_i, s_{i-1} and u_{i-1} (see previous section), we create additional training instances keeping the target system utterance s_i and previous system utterance s_{i-1} the same, but using a paraphrase u'_i in the place of u_i, and a paraphrase u'_{i-1} in the place of u_{i-1}. Training instances are added for all possible combinations of the available paraphrases for u_i and u_{i-1}, providing some (artificial) coverage for parts of the space of possible dialogue paths that would be otherwise completely ignored during training.

Training the classifier with material from the external referees (see previous section) and additional synthetic training examples from paraphrases as described above produces a dialogue policy with weak accuracy of 66%, at the same level as the policy learned with manually assigned speech acts. It is noteworthy that this was achieved through a very simple and intuitive paraphrase annotation task that requires no technical knowledge about dialogue systems, dialogue acts or domain modeling. As mentioned in section 3.2, paraphrases for each of the 19 dialogues were generated in less than 30 minutes on average.

6 Conclusion and future work

We introduced a framework for collection and enrichment of scenario-specific dialogues based only on tasks that require no technical knowledge. Data collected in this framework support novel approaches not just for learning dialogue policies, but perhaps more importantly for evaluating learned policies, which allows us to examine different techniques using an objective automatic metric.

Although research on both learning and evaluating dialogue policies is still in early stages, this case study and proof-of-concept experiments serve to illustrate the basic ideas of external referee and paraphrase annotation, and the use of multiple reference dialogue act choices in evaluation of dialogue policies, in a way similar to how multiple reference translations are used in evaluation of machine translation systems. We do not consider this line of research a replacement for or an alternative to formal modeling of domains and dialogue behavior, but rather as an additional tool in the community's collective arsenal. There are many unexplored avenues for including data-driven techniques within rule-based frameworks and vice-versa.

In future work we intend to further validate the ideas presented in this paper by performing additional collection of dialogues in the Amani domain to serve as a virgin test set, and applying these techniques to other dialogue domains and scenarios. We also plan to refine the weak accuracy and weak agreement metrics to take into account the level of agreement within utterances to reflect that some parts of dialogues may be more open-ended than others. Finally, we will conduct human evaluations of different policies to begin validating weak accuracy as an automatic metric for evaluation of dialogue policies.

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. We would also like to thank Ron Artstein, Sudeep Gandhe, Fabrizio Morbini, Angela Nazarian, Susan Robinson, Michael Rushforth, and David Traum.

References

S. Abney, S. Flickenger, C. Gdaniec, C. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski. 1991. Procedure for quantitatively comparing the syntactic coverage of english grammars. In E. Black, editor, *Proceedings of the workshop on Speech and Natural Language*, HLT '91, pages 306–311, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Sudeep Gandhe and David R. Traum. 2007. Creating spoken dialogue characters from corpora without annotations. In *Proceedings of Interspeech-07*, 08/2007.

Sudeep Gandhe and David R. Traum. 2010. I've said it before, and i'll say it again: An empirical investigation of the upper bound of the selection approach to dialogue. In *11th 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.

Sina Jafarpour, Chris Burges, and Alan Ritter. 2009. Filter, rank, and transfer the knowledge: Learning to chat. In *Proceedings of the NIPS Workshop on Advances in Ranking*.

Patrick Kenny, Thomas D. Parsons, Jonathan Gratch, Anton Leuski, and Albert A. Rizzo. 2007. Virtual patients for clinical therapist skills training. In *Proceedings of the 7th international conference on Intelligent Virtual Agents*, IVA '07, pages 197–210, Berlin, Heidelberg. Springer-Verlag.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Proceedings of HLT-NAACL*.

Linus Sellberg and Arne Jnsson. 2008. Using random indexing to improve singular value decomposition for latent semantic analysis. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may.

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.

Ingrid Zukerman and Yuval Marom. 2006. A corpus-based approach to help-desk response generation. *Computational Intelligence for Modelling, Control and Automation, International Conference on*, 1:23.