ADiEU: Toward Domain-Based Evaluation of Spoken Dialog Systems

Jan Kleindienst, Jan Cuřín, and Martin Labský

IBM Research, Prague, Czech Republic
{jankle,jan_curin,martin.labsky}@cz.ibm.com

Abstract. We propose a new approach toward evaluation of spoken dialog systems. The novelty of our method is based on utilization of domain-specific knowledge combined with the deterministic measurement of dialog system performance on a set of individual tasks within the domain. The proposed methodology thus attempts to answer questions such as: "How well is my dialog system performing on a specific domain?", "How much has my dialog system improved since the previous version?", "How much is my dialog system better/worse than other dialog systems performing on that domain?"

Keywords: Dialog, evaluation, scoring, multimodal, speech recognition.

1 Introduction

Research in the field of conversational and dialog systems has a long tradition starting in 1966 with Weizenbaum's Eliza [1]. More recently, research in spoken dialog systems has tackled more ambitious domains, such as problem solving [2], navigation [3], or tutoring systems [4].

This paper is organized as follows. In introduction we outline our motivation and the principle of the proposed method. In Section 2 we introduce the concept of a domain task ontology that can serve as a benchmarking tool for well-known application domains. Section 3 describes in detail the proposed ADiEU metric and its computation. Section 4 presents a case study in the music management domain and demonstrates the application of ADiEU to a real-world task. We discuss practical considerations regarding the proposed metric in Section 5, human evaluation in Section 6, and conclude in Section 7.

1.1 Rationale

Current methods and techniques for measuring performance of spoken dialog systems are still very immature. They are either based on subjective evaluation (Wizard of Oz or other usability studies) or they are borrowing automatic measures used in speech recognition, machine translation or action classification, which provide only incomplete picture of the performance of the system. Nowadays, dialog systems are evaluated by measures used in speech recognition, such as word error rate (WER) or action classification error rate [5], by techniques that measure primarily dialog coherence [6], and by systems supporting human judgment-based evaluation, such as PARADISE [7, 8]. What

J.A. Jacko (Ed.): Human-Computer Interaction, Part I, HCII 2009, LNCS 5610, pp. 287–294, 2009.

is particularly missing in this area is (1) a measurement of performance for a particular domain, (2) possibility to compare one dialog system with others, and (3) evaluation of a progress during the development of a dialog system. By the ADiEU[1] scoring presented herein we attempt to address these three cases.

1.2 The Elements of ADiEU Metric

The ADiEU score consists of two ingredients both of which range from 0 to 1:

A) Domain Coverage (DC) score,
B) Dialog Efficiency (DE) score.

We describe both scores in the following chapters. Note that the results of domain coverage and dialog efficiency may be combined into a single compound score to attain a single overall characteristic (the eigen value) of the assessed dialog system.

The ADiEU score relies on a good understanding of the dialog domain that is described in the form of a *domain task ontology*. The more expert knowledge is projected into the domain ontology, the more reliable results we expect from the ADiEU score.

2 Capturing Domain Ontology

The cornerstone of our approach is to evaluate spoken and multi-modal dialog systems within a predefined, well-known (and typically narrow) domain. In our labs we have developed many speech and multimodal applications for various domains, such as music selection, TV remote control, in-car navigation and phone control; using grammars, language models and natural language understanding techniques. In order to compare two spoken dialog systems that deal with the same domain, we first describe the domain diligently using the task ontology. This restricted ontology represents the human expert knowledge of the domain and is encoded as a set of tasks with two kinds of relations between the tasks: task generalization and aggregation. Individual tasks are defined as sequences of parameterized actions. Actions are separable units of domain functionality, such as volume control, song browsing or playback.

Parameters are categories of named entities, such as album or track title, artist name or genre. Tasks are labeled by weights, which express the relative importance of a particular task with respect to other tasks. The ontology may also define task aggregations which explicitly state that a complex task can be realized by sequencing several simpler tasks. Table 1 shows a sample task ontology for the music control domain. For example, the task volume control/relative with weight of 2 (e.g. "louder, please") is considered more important in evaluation than its absolute sibling (e.g. "set volume to 5"). This may be highly subjective if scored by a single human judge and thus a consensus of domain experts may be required to converge to a generally acceptable ontology for the domain. Once acknowledged by the community, this ontology could be used as the common etalon for scoring third-party dialog systems.

[1] We call our measurement the Automatic Dialog Evaluation Understudy, ADiEU.

Table 1. Speech-enabled reference tasks for the jukebox domain. Tasks are divided into groups. Both group as well as tasks within the group are assigned relative importance points by an expert. These points are normalized to obtain per-task contribution to the domain's functionality. ITC shows ideal turn count range for each task.

Group		Task			
Points	Share	Description	Points	Contrib. %	ITC
Volume		relative	2	6.20	1
2	15.50%	absolute	1	3.10	1
		mute	2	6.20	1
Playback		play	3	7.75	1
4	31.01%	stop	3	7.75	1
		pause	1.5	3.88	1
		resume	1.5	3.88	1
		next, previous track	1	2.58	1
		next, previous album	1	2.58	1
		media selection	1	2.58	1
Play mode		shuffle	1	1.94	1
0.5	3.88%	repeat	1	1.94	1
Media library		browse by criteria	2	3.93	1..2
6	46.51%	play by criteria	4	7.85	1..2
		search by genre	2	3.93	1
		search by artist name			
		up to 100 artists	1	1.96	1..2
		more then 100 artists	2	3.93	1..2
		search by album name			
		up to 200 albums	1	1.96	1..2
		more than 200 albums	2	3.93	1..2
		search by song title			
		up to 250 songs	1	1.96	1..2
		more than 2000 songs	2	3.93	1..2
		search by partial names			
		words	1	1.96	2
		spelled letters	1	1.96	2
		ambiguous entries	2	3.93	2
		query			
		item counts	0.5	0.98	1
		favorites			
		browse and play	0.5	0.98	1..2
		add items	0.3	0.59	1
		media management			
		refresh from media	0.2	0.39	1
		add or remove media	0.2	0.39	1..2
		access online content	1	1.96	2..3
Menu		quit	0.5	1.03	1..2
0.4	3.10%	switch among other apps	1	2.07	1..2
	100%			100	

3 The Proposed Method of ADiEU Evaluation

The actual dialog system evaluation metric that is in the heart of our method consists of two indicators: Domain Coverage (DC) - computed over the task ontology and Dialog Efficiency (DE) that quantifies the outcome of user test sessions. The DC expresses how the evaluated system covers the set of tasks in the ontology for a particular domain; while the DE indicates the performance of the evaluated system on those tasks supported by the system.

3.1 Scoring of Domain Coverage

The domain coverage (*DC*) is a sum of weights of tasks supported by the system (*S*) over the sum of weights of all tasks from the ontology (*O*).

$$DC(S,O) = \frac{\sum_{t \in supported\ tasks(O)} w_t}{\sum_{t^* \in all\ tasks(O)} w_{t^*}} \tag{1}$$

Table 1 shows a sample domain task ontology for the music management domain that shows the raw points assigned by a domain expert and their normalized versions that are used to assess the relative importance of individual tasks. The expert may control the weights of whole task groups (such as *Playback control*) as well as the weights of individual tasks that comprise these groups. Generally, the ontology can have more than two levels of sub-categorization that are shown in the example.

3.2 Scoring of Dialog Efficiency

The actual efficiency of dialog is measured using the number of dialogue turns [9, 10] needed to accomplish a chosen task. In spoken dialog systems, a dialog turn corresponds to a pattern of user speech input followed by the system's response. We introduce a generalized *penalty turn count* (PTC) that measures overall dialog efficiency by incorporating other considered factors: number of help requests, number of rejections, and user and system reaction times.

$$PTC(t) = TC(t) + \lambda_{NHR}NHR(t) + \lambda_{NRP}NRP(t) + \lambda_{URT}URT(t) + \lambda_{SRT}SRT(t) \tag{2}$$

Where TC is the actual dialog turn count, NHR is the number of help requests, URT is user response time and SRT is system response time and the lambdas represent weights of each contributor to the final penalty turn count (PTC)[2].

The obtained penalty turn count in then compared to an ideal number of turns for a particular task. We define a key property, the *ideal number of turns* (INT), as being determined by at least the following factors. The INT is (F1) directly proportional to a number of information slots to be filled and (F2) indirectly proportional to a size of the block of information slots commonly accepted as coherent.

$$INT(t) = \frac{number\ of\ information\ slots\ to\ be\ filled}{size\ of\ a\ block\ of\ information\ slots\ commonly\ accepted\ as\ coherent} \tag{3}$$

For example, the concept of “date” consists of three information slots (day, month, and year) that need to be filled. Here, the number of information slots (F1) is three, which is in this case the same as the size of a coherent block expected by the users. The INT for the “date” concept is thus 1 (=3/3). In the contemporary art the INT property is determined manually by a human judgment.

[2] In our experiments, we set λ_{NHR}=0.5, λ_{NRP}=1, and λ_{URT}=λ_{SRT}=0 since for the music domain the user reaction time was not indicative of dialog quality and both applications responded instantly.

The actual score of the dialog efficiency (DE score) for an individual task is then counted as a fraction of difference between INT and PTC against current PTC, i.e.:

$$DE(t) = 1 - \max\left(\frac{PTC(t) - INT(t)}{PTC(t)}, 0 \right) \tag{4}$$

To avoid subjective scoring we typically use several human testers as well as several trials per one task. For example for task “play by artist” the following set of trials can be used: “Play something by Patsy Cline”, “Play some song from your favorite interpreter”, or “Play some rock album, make the final selection by the artist name”. Each of these trials has assigned its ideal number of turns (this is why INT for tasks in the ontology are given by range in the Table 1.) The task dialog efficiency score is then computed as an average over all human testers and dialog efficiency for each trial. Samples of trials used in the evaluation of music management domain are given in Table 2.

3.3 The ADiEU Score

The ADiEU score is then counted as a sum of products of domain coverage and dialog efficiency for each task in the domain ontology, i.e.:

$$ADiEU(S,O) = \frac{\sum_{t \in supported\ tasks(O)} w_t \cdot DE(t)}{\sum_{t \in supported\ tasks(O)} w_t} \tag{5}$$

4 Case Study: ADiEU Scores for Music Management Domain

We applied the ADiEU scoring to our two dialog systems developed at different times and both partially covering the music management dialog domain. Both allow their users to play music by dynamically generating grammars based on meta tags found in users’ mp3 files. The first one, named *A-player*, is simpler and covers a limited part of the music management domain. The second, named *Jukebox*, covers a larger part of the domain and also allows free-form input using a combination of statistical language models and maximum entropy based action classifiers.

For both applications, we collected input from a group of 10 speakers who were asked to accomplish tasks listed in Table 2. Each of these user tasks corresponded to a task in the domain task ontology and there was at least one user task per each ontology task that was supported by either A-player or Jukebox. The subjects were given general guidance but no sample English phrases were suggested to them that could be used to control the system. In order not to guide users even by the wording of the user tasks, the tasks were described to them in their native language. All ten subjects were non-native but fluent English speakers.

Table 2. Specific tasks to be accomplished by speakers using A-player and Jukebox

Task	A-player	Jukebox	ITC
Start playback of arbitrary music	x	x	1
Increase the volume		x	1
Set volume to level 10		x	1
Mute on		x	1
Mute off		x	1
Pause		x	1
Resume		x	1
Next track	x	x	1
Previous track	x	x	1
Shuffle	x	x	1
Play some jazz song		x	1
Play a song from Patsy Cline	x	x	1
Play Iron Man from Black Sabbath	x	x	1
Play the album The Best of Beethoven	x	x	1
Play a song Where the Streets Have No Name	x	x	1
Play a song Sonata no. 11 (ambiguous)	x	x	2
Play a rock song by your favorite artist	x	x	3
Reload songs from media	x		1

Table 3. Computation of coverage, task completion score and ADiEU for A-player and Jukebox

Task	A-player				Jukebox			
	sup	DC weight	DE score	ADiEU	sup	DC weight	DE score	ADiEU
volume relative	0	0		0.000	1	6.2	0.82	0.051
volume absolute	0	0		0.000	1	3.1	0.82	0.025
mute	0	0		0.000	1	6.2	0.82	0.051
play	1	7.75	0.57	0.044	1	7.75	0.32	0.025
stop	1	7.75	1.00	0.078	1	7.75	0.82	0.064
pause	0	0		0.000	1	3.88	0.57	0.022
resume	0	0		0.000	1	3.88	0.50	0.019
next, prev. track	1	2.58	0.80	0.021	1	2.58	1.00	0.026
next, prev. album	0	0	0.50	0.000	1	2.58	0.80	0.021
shuffle	0	0		0.000	1	1.94	0.67	0.013
browse by criteria	0	0		0.000	0.5	1.97	0.52	0.010
play by criteria	1	7.85	0.82	0.064	1	7.85	0.67	0.052
search by genre	0	0	0.67	0.000	1	3.93	0.78	0.030
search by artist				0.000				0.000
<= 100 artists	1	1.96	0.83	0.016	1	1.96	0.40	0.008
> 100 artists	1	3.93	0.83	0.033	1	3.93	0.60	0.024
search by album				0.000				0.000
<= 200 albums	1	1.96	1.00	0.020	1	1.96	0.29	0.006
> 200 albums	1	3.93	1.00	0.039	1	3.93	0.75	0.029
search by song				0.000				0.000
<= 250 songs	1	1.96	0.79	0.015	1	1.96	0.61	0.012
> 2000 songs	1	3.93	0.79	0.031	1	3.93	0.93	0.036
word part. search	0	0		0.000	1	1.96	0.55	0.011
ambiguous entries	0	0		0.000	1	3.93	0.49	0.019
media refresh	1	0.39	0.67	0.003	0	0		0.000
	0.34	**43.99**	**82.6**	**36.3**	0.55	**83.17**	**66.7**	**0.554**

Table 3 shows the computation of the ADiEU score and its components: *domain coverage* (DC) and *domain efficiency* (DE). For A-player, which is limited in functionality, the weighted *domain coverage* only reached 43.99%, whereas for Jukebox this was 83.17%. On the other hand, A-player allowed its users to accomplish the tasks it supported more quickly than Jukebox; this is documented by the weighted *dialog efficiency* score reaching 82.6% for A-player and 66.7% for Jukebox. This was mainly due to Jukebox being more interactive (e.g. asking questions, presenting choices) and due to a slightly higher error rate of a dictation-based system as opposed to a grammar-based one. The overall ADiEU score was higher for Jukebox (55.4%) than it was for A-player (36.3%). This was in accord with the feedback we received from users from ongoing evaluations who claimed they had better experience with the Jukebox application. The two major reasons were the support of free-form commands by the Jukebox and its broader functionality.

5 Human Evaluation in Progress

The HCI methodology [10] advocates several factors that human judges collect in the process of dialog system evaluation. These key indicators include accuracy, intuitiveness, reaction time, and efficiency. When designing the evaluation method we attempted to incorporate the core of these indicators into the scoring method to ensure good correlation of the ADiEU metric with the human judgment. We are currently collecting data form the evaluation test where the human judges act as personas [11]. The results of the evaluation either confirm or reject the assumption of the ADiEU scoring correlation with human judgment.

6 Practical Considerations of the ADiEU Scoring

The application of the ADiEU scoring to an arbitrary dialog system has several practical considerations. Generally, there are two possibilities how to evaluate a third-party dialog system by our metric: 1) agreed API contract supported by the external system or 2) rich enough tracing and logging information. Both approaches will typically require cooperation with the supplier of the measured system. The API approach asserts there exists a runtime API that supports e.g.: simulating input to the system, changing the dialog state, obtaining notification about dialog state changes with sufficient introspection, possibility to read output of the system. The logging approach demands the application to write all the required information to a log file, ideally in a format compliant with the ADiEU score measuring tool. This usually means tight cooperation with the dialog system engineers, but it is easier and more straight forward than changing the application API in the case it does not provide access to all information needed by the ADiEU metric. Having the test run in the form of log has the advantage of the possibility to send the logs to the scoring tool hosted as a web service and the possibility to evaluate the system against multiple domain ontologies or ontology versions of the same domain. We have experimented with both approaches while evaluation our systems.

7 Conclusion

We introduce a method for quantitative evaluation of spoken dialog system that utilizes the domain knowledge encoded by a human expert. The evaluation results are described in the form of a comparison metric consisting of domain coverage and dialog efficiency scores allowing to compare relative as well as absolute performance of a system within a given domain. This approach has an advantage of comparing incremental improvements on an individual dialog system that the dialog designer may want to verify along the way. In addition, the method allows to cross-check the performance of third-party dialog systems operating on the same domain and immediately understand the strong and weak points in the dialog design. Human evaluations are currently conducted to estimate the correlation between the ADiEU score and human judgment. The subjectivity of human scoring and consensus on the ontology coverage are subject of further investigation.

References

1. Weizenbaum, J.: ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine. Communications of the Association for Computing Machinery 9, 36–45 (1966)
2. Allen, J., Chambers, N., Ferguson, G., Galescu, L., Jung, H., Swift, M., Taysom, W.: PLOW: A Collaborative Task Learning Agent. In: Twenty-Second Conference on Artificial Intelligence, AAAI-2007 (2007)
3. Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K.: Mack: Media lab autonomous conversational kiosk. In: Imagina 2002 (2002)
4. Graesser, A.C., VanLehn, K., Rosfie, C.P., Jordan, P.W., Harter, D.: Intelligent tutoring systems with conversational dialogue. AI Mag. 22(4), 39–51 (2001)
5. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (International edn.). Prentice-Hall, Englewood Cliffs (February 2000)
6. Gandhe, S., Traum, D.: Evaluation understudy for dialogue coherence models. In: Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, Columbus, Ohio, June 2008, pp. 172–181. Association for Computational Linguistics (2008)
7. Walker, M., Kamm, C., Litman, D.: Towards developing general models of usability with paradise. Nat. Lang. Eng. 6(3-4), 363–377 (2000)
8. Hajdinjak, M., Mihelific, F.: The paradise evaluation framework: Issues and findings. Comput. Linguist. 32(2), 263–272 (2006)
9. Le Bigot, L., Bretier, P., Terrier, P.: Detecting and exploiting user familiarity in natural language human-computer dialogue. In: Asai, K. (ed.) Human Computer Interaction: New Developments, pp. 269–382. InTech Education and Publishing (2008); ISBN: 978-953-7619-14-5
10. Nielsen, J.: Heuristic evaluation. In: Nielsen, J., Mack, R.L. (eds.) Usability Inspection Methods, pp. 25–64. John Wiley & Sons, New York (1994); ISBN: 0-471-01877-5
11. Carroll, J.: Human Computer Interaction in the New Millennium. ACM Press, New York (2001)