I've said it before, and I'll say it again: An empirical investigation of the upper bound of the selection approach to dialogue

Sudeep Gandhe and **David Traum**
Institute for Creative Technologies
13274 Fiji way, Marina del Rey, CA 90292
{gandhe,traum}@ict.usc.edu

Abstract

We perform a study of existing dialogue corpora to establish the theoretical maximum performance of the selection approach to simulating human dialogue behavior in unseen dialogues. This maximum is the proportion of test utterances for which an exact or approximate match exists in the corresponding training corpus. The results indicate that some domains seem quite suitable for a corpus-based selection approach, with over half of the test utterances having been seen before in the corpus, while other domains show much more novelty compared to previous dialogues.

1 Introduction

There are two main approaches toward automatically producing dialogue utterances. One is the *selection* approach, in which the task is to pick the appropriate output from a corpus of possible outputs. The other is the *generation* approach, in which the output is dynamically assembled using some composition procedure, e.g. grammar rules used to convert information from semantic representations and/or context to text.

The generation approach has the advantage of a more compact representation for a given generative capacity. But for any finite set of sentences produced, the selection approach could perfectly simulate the generation approach. The generation approach generally requires more analytical effort to devise a good set of grammar rules that cover the range of desired sentences but do not admit undesirable or unnatural sentences. Whereas, in the selection approach, outputs can be limited to those that have been observed in human speech. This affords complex and human-like sentences without much detailed analysis. Moreover, when the output is not just text but presented as speech, the system may easily use recorded audio clips rather than speech synthesis. This argument also extends to multi-modal performances, e.g. using artist animation motion capture or recorded video for animating virtual human dialogue characters. Often one is willing to sacrifice some generality in order to achieve more human-like behavior than is currently possible from generation approaches.

The selection approach has been used for a number of dialogue agents, including question-answering characters at ICT (Leuski et al., 2006; Artstein et al., 2009; Kenny et al., 2007), FAQ bots (Zukerman and Marom, 2006; Sellberg and Jönsson, 2008) and web-site information characters. It is also possible to use the selection approach as a part of the process, e.g. from words to a semantic representation or from a semantic representation to words, while using other approaches for other parts of dialogue processing.

The selection approach presents two challenges for finding an appropriate utterance:

- *Is there a good enough utterance to select?*
- *How good is the selection algorithm at finding this utterance?*

We have previously attempted to address the second question, by proposing the information ordering task for evaluating dialogue coherence (Gandhe and Traum, 2008). Here we try to address the first question, which would provide a theoretical upper bound in quality for any selection approach. We examine a number of different dialogue corpora as to their suitability for the selection approach.

We make the following assumptions to allow automatic evaluation across a range of corpora. Actual human dialogues represent a gold-standard for computer systems to emulate; i.e. choosing an actual utterance in the correct place is the best possible result. Other utterances can be evaluated as to how close they come to the original utterance,

Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 245–248,
The University of Tokyo, September 24-25, 2010.

using a similarity metric.

Our methodology is to examine a test corpus of human dialogue utterances to see how well a selection approach could approximate these, given a training corpus of utterances in that domain. We look at exact matches as well as utterances having their similarity score above a threshold. We investigate the effect of the size of training corpora, which lets us know how much data we might need to achieve a certain level of performance. We also investigate the effect of domain of training corpora.

2 Dialogue Corpora

We examine human dialogue utterances from a variety of domains. Our initial set contains six dialogue corpora from ICT as well as three other publicly available corpora.

SGT Blackwell is a question-answering character who answers questions about the U.S. Army, himself, and his technology. The corpus consists of visitors interacting with SGT Blackwell at an exhibition booth at a museum. **SGT Star** is a question-answering character, like SGT Blackwell, who talks about careers in the U.S. Army. The corpus consists of trained handlers presenting the system. **Amani** is a bargaining character used as a prototype for training soldiers to perform tactical questioning. The **SASO** system is a negotiation training prototype in which two virtual characters negotiate with a human "trainee" about moving a medical clinic. The **Radiobots** system is a training prototype that responds to military calls for artillery fire. **IOTA** is an extension of the Radiobots system. The corpus consists of training sessions between a human trainee and a human instructor on a variety of missions. Yao et al. (2010) provides details about the ICT corpora.

Other corpora involved dialogues between two people playing specific roles in planning, scheduling problem for railroad transportation, the **Trains-93** corpus (Heeman and Allen, 1994) and for emergency services, the **Monroe** corpus (Stent, 2000). The **Switchboard** corpus (Godfrey et al., 1992) consists of telephone conversations between two people, based on provided topics.

We divided the data from each corpus into a training set and a test set, as shown in Table 1. The data consists of utterances from one or more human speakers who engage in dialogue with either virtual characters (Radiobots, Blackwell, Amani, Star, SASO) or other humans (Switchboard, Monroe, IOTA, Trains-93). These corpora differ along a number of dimensions such as the size of the corpus, dialogue genre (question-answering, task-oriented or conversational), types of tasks (artillery calls, moving and scheduling resources, information seeking) and motivation of the participants (exploring a new technology – SGT Blackwell , presenting a demo – SGT Star, undergoing training – Amani, IOTA or simply for collecting the corpus – Switchboard, Trains-93, Monroe). While the set of corpora we include does not cover all points in these dimensions, it does present an interesting range.

3 Dialogue Utterance Similarity Metrics

To answer the question of whether an adequate utterance exists in our training corpus that could be selected and used, we need an appropriateness measure. We assume that an utterance produced by a human in a dialogue is appropriate, and thus the problem becomes one of constructing an appropriate similarity function to compare the human-produced utterance with the utterances available from the training corpus. Given a training corpus U_{train} and a similarity function f , we calculate the score for a test utterance u_t as, $maxsim_f(u_t) = \max_i f(u_t, u_i); u_i \in U_{train}$ There are several choices for the utterance similarity function f . Ideally such a function would take meaning and context into account rather than just surface similarity, but these aspects are harder to automate, so for our initial experiments we look at several surface metrics, as described below.

Exact measure returns 1 if the utterances are exactly same and 0 otherwise. **1-WER**, a similarity measure related to word error rate, is defined as $min\,(0, 1 - levenshtein(u_t, u_i)/length(u_t))$. **METEOR** (Lavie and Denkowski, 2009), one of the automatic evaluation metrics used in machine translation is a good candidate for f. METEOR finds optimal word-to-word alignment between test and reference strings based on several modules that match exact words, stemmed words and synonyms. METEOR is a tunable metric and for our analysis we used the default parameters tuned for the Adequacy & Fluency task. All previous measures take into account the word ordering of test and reference strings. In contrast, document similarity measures used in information retrieval generally follow the *bag of words* assumption, where a

Domain	Train		Test		$mean(maxsim_f)$				% of utterances		
										METEOR	
	# utt	words	# utt	words	METEOR	1-WER	Dice	Cosine	Exact	≥ 0.9	≥ 0.8
Blackwell	17755	84.7k	2500	12.0k	0.913	0.878	0.917	0.921	69.6	75.8	82.1
Radiobots	995	6.8k	155	1.2k	0.905	0.864	0.920	0.924	53.6	67.7	83.2
SGT Star	2974	16.6k	400	2.2k	0.897	0.860	0.906	0.911	65.0	70.5	78.0
SASO	3602	23.3k	510	3.6k	0.821	0.742	0.830	0.837	38.4	48.6	62.6
IOTA	4935	50.4k	650	5.6k	0.768	0.697	0.800	0.808	36.2	42.8	51.4
Trains 93	5554	47.2k	745	6.0k	0.729	0.633	0.758	0.769	34.5	36.9	42.8
SWBD[1]	19741	138.2k	3173	21.5k	0.716	0.628	0.736	0.753	35.8	37.9	44.2
Amani	1455	15.8k	182	1.9k	0.675	0.562	0.694	0.706	18.7	25.8	30.8
Monroe	5765	43.0k	917	8.8k	0.594	0.491	0.639	0.658	22.3	23.6	26.1

Table 1: Corpus details and within domain results

string is converted to a set of tokens. Here we also considered **Cosine** and **Dice** coefficients using the standard boolean model. In our experiments, the surface text was normalized and all punctuation was removed.

4 Experiments

Results Within a Domain

In our first experiment, we computed $maxsim_f$ scores for all test corpus utterances in a given domain using the training utterances from the same domain. For the domains Blackwell, SGT Star, SASO, Amani & Radiobots which are implemented dialogue systems our corpus consists of user utterances only. For Trains 93 and Monroe corpora, we make sure to match the speaker roles for u_t and u_i. For Switchboard, where speakers do not have any special roles and for IOTA, where the speaker information was not readily accessible, we ignore the speaker information and select utterances from either speaker.

Table 1 reports the mean of $maxsim_f$ scores. These can be interpreted as the expectation of $maxsim_f$ score for a new test utterance. The higher this expectation, the more likely it is that an utterance similar to the new one has been seen before and thus the domain will be more amenable to selection approaches. This table also shows the percentage of utterances that had a $maxsim_{Meteor}$ score above a certain threshold. The correlation between $maxsim_f$ for different choices of f (except Exact match) is very high (Pearson's $r > 0.94$). The histogram analysis shows that SGT Star, Blackwell, Radiobots and SASO domains are better suited for *selection* approaches. Domains like Trains-93, Monroe, Switchboard and Amani have a more diffuse distribution and are not best suited for *selection* approaches, at least with the amount of data we have available. The IOTA domain falls somewhere in between these two domain classes.

[1]Switchboard (SWBD) is a very large corpus and for running our experiments in a reasonable computing time we only selected a small portion of it.

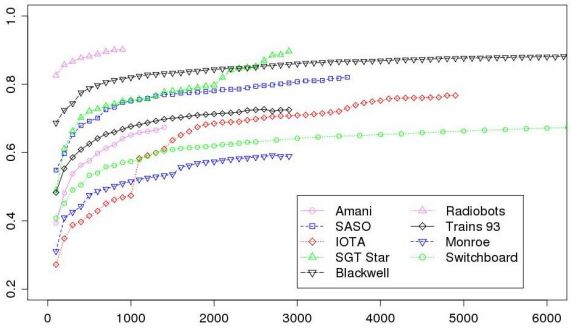

Figure 1: $maxsim_{Meteor}$ vs # utterances in training data for different domains

Effect of Training Data Size

Figure 1 shows the effect of training data size on the $maxsim_{Meteor}$ score. Radiobots shows very high scores even for small amounts of training data. SGT Star and SGT Blackwell also converge fairly early. Switchboard, on the other hand, does not achieve very high scores even with a large number of utterances. For all domains, with around 2500 training utterances $maxsim_{Meteor}$ reaches 90% of its maximum possible value for the training set.

Comparing Different Domains

In order to understand the similarities between different dialogue domains, we computed $maxsim_{Meteor}$ for a test domain using training

		Training Domains								
		IOTA	Radio-bots	SGT Star	Black-well	Amani	SASO	Trains-93	Monroe	SWBD
Testing Domains	IOTA	**0.768**	0.440	0.247	0.334	0.196	0.242	0.255	0.297	0.334
	Radiobots	0.842	**0.905**	0.216	0.259	0.161	0.183	0.222	0.270	0.284
	SGT Star	0.324	0.136	**0.897**	0.622	0.372	0.438	0.339	0.417	0.527
	Blackwell	0.443	0.124	0.671	**0.913**	0.507	0.614	0.424	0.534	0.696
	Amani	0.393	0.134	0.390	0.561	**0.675**	0.478	0.389	0.420	0.509
	SASO	0.390	0.125	0.341	0.516	0.459	**0.821**	0.443	0.454	0.541
	Trains 93	0.434	0.112	0.214	0.468	0.272	0.429	**0.753**	0.627	0.557
	Monroe	0.409	0.119	0.217	0.428	0.276	0.404	0.534	**0.630**	0.557
	SWBD	0.368	0.110	0.280	0.490	0.362	0.383	0.562	0.599	**0.716**

Table 2: Mean of $maxsim_{Meteor}$ for comparing different dialogue domains. The **bold-faced** values are the highest in the corresponding row.

sets from other domains. In this exercise, we ignored the speaker information. Table 2 reports the mean values of $maxsim_{Meteor}$ for different training domains. For all the testing domains, using the training corpus from the same domain produces the best results. Notice that Radiobots also has good performance with the IOTA training data. This is as expected since IOTA is an extension of Radiobots and should cover a lot of utterances from the Radiobots domain. Switchboard and Blackwell training corpora have a overall higher score for all testing domains. This may be due to the breadth and size of these corpora. On the other extreme, the Radiobots training domain performs very poorly on all testing domains other than itself.

5 Discussion

We have examined how well suited a corpus-based selection approach to dialogue can succeed at mimicking human dialogue performance across a range of domains. The results show that such an approach has the potential of doing quite well for some domains, but much less well for others. Results also show that for some domains, quite modest amounts of training data are needed for this operation. Applying this method across corpora from different domains can also give us a similarity metric for dialogue domains. Our hope is that this kind of analysis can help inform the decision of what kind of language processing methods and dialogue architectures are most appropriate for building a dialogue system for a new domain, particularly one in which the system is to act like a human.

Acknowledgments

This work 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 like to thank Ron Artstein and others at ICT for compiling the ICT Corpora used in this study.

References

R. Artstein, S. Gandhe, J. Gerten, A. Leuski, and D. Traum. 2009. Semi-formal evaluation of conversational characters. In *Languages: From Formal to Natural. Essays Dedicated to Nissim Francez on the Occasion of His 65th Birthday*, volume 5533 of *LNCS*. Springer.

S. Gandhe and D. Traum. 2008. Evaluation understudy for dialogue coherence models. In *Proc. of SIGdial 08.*

J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proc. of ICASSP-92*, pages 517–520.

P. A. Heeman and J. Allen. 1994. The TRAINS 93 dialogues. TRAINS Technical Note 94-2, Department of Computer Science, University of Rochester.

P. Kenny, T. Parsons, J. Gratch, A. Leuski, and A. Rizzo. 2007. Virtual patients for clinical therapist skills training. In *Proc. of IVA 07*, Paris, France. Springer.

A. Lavie and M. J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23:105–115.

A. Leuski, R. Patel, D. Traum, and B. Kennedy. 2006. Building effective question answering characters. In *Proc. of SIGdial 06*, pages 18–27, Sydney, Australia.

L. Sellberg and A. Jönsson. 2008. Using random indexing to improve singular value decomposition for latent semantic analysis. In *Proc. of LREC'08*, Morocco.

A. J. Stent. 2000. The monroe corpus. Technical Report 728, Computer Science Dept. University of Rochester.

X. Yao, P. Bhutada, K. Georgila, K. Sagae, R. Artstein, and D. Traum. 2010. Practical evaluation of speech recognizers for virtual human dialogue systems. In *LREC 2010.*

I. Zukerman and Y. Marom. 2006. A corpus-based approach to help-desk response generation. In *CIMCA/IAWTIC '06.*