Creating Spoken Dialogue Characters from Corpora without Annotations

Sudeep Gandhe, David Traum

Institute for Creative Technologies, University of Southern California,
13274 Fiji way, Marina del Rey, CA, 90292, USA
gandhe@ict.usc.edu, traum@ict.usc.edu

Abstract

Virtual humans are being used in a number of applications, including simulation-based training, multi-player games, and museum kiosks. Natural language dialogue capabilities are an essential part of their human-like persona. These dialogue systems have a goal of being believable and generally have to operate within the bounds of their restricted domains. Most dialogue systems operate on a dialogue-act level and require extensive annotation efforts. Semantic annotation and rule authoring have long been known as bottlenecks for developing dialogue systems for new domains. In this paper, we investigate several dialogue models for virtual humans that are trained on an unannotated human-human corpus. These are inspired by information retrieval and work on the surface text level. We evaluate these in text-based and spoken interactions and also against the upper baseline of human-human dialogues.

Index Terms: virtual humans, dialogue modelling, human-human corpus

1. Introduction

Virtual human characters have proved useful in many fields, including simulation, training and interactive games. An ability to take part in conversations using natural language is important for believable virtual humans. This interface has to be good enough to engage the trainee or the gamer in the activity.

Natural language dialogue systems come in many different flavors. Chatterbot systems like Eliza [1] or Alice [2] have to operate in an unrestricted domain with an aim of being human-like. On the other hand, task-oriented dialogue systems such as pizza-ordering, ATIS [3] or Trains [4] restrict the user quite severely in the allowed topics and ways of talking about them.

Fig 1 summarizes some different types of dialogue systems and how virtual humans compare to them. For chatbots, it is often sufficient to talk about topics at a fairly shallow level, without requiring a lot of detailed task knowledge or knowledge of how some parts of a task relate to others. This luxury is not available for a task oriented dialogue where the system is expected to perform a task or provide task-specific information.

There are some domains that fall between these extremes, for instance negotiation about whether or not to adopt a proposal. In this case, there is definitely a task or set of tasks involved, but one does not necessarily require as detailed knowledge as is required to actually perform the task.

There are also various methods for dialogue management. Chatbots typically follow Eliza in operating at a textual level, with pattern matching. Some methods like corpus-based retrieval approaches (e.g., [5, 6]) have an advantage of robust selection, with a more limited set of responses. Other methods include trying to learn the pattern matching rules from a corpus [7]. Task oriented dialogue system generally operates at a concept or dialogue act level allowing for easy integration with other kinds of knowledge-based reasoning, but at the price of more processing to translate from the surface level to the higher level of abstraction and back. All of these methods require either extensive writing of rules or other symbolic processing methods, or extensive corpus annotations, both of which serve to introduce a high cost in the development of a dialogue system for a new domain.

In our previous work [8] we implemented unsupervised corpus based methods to bootstrap dialogue bots. These bots don't have sophisticated cognitive models, but they can be built instantly from a human human dialogue corpus without annotation or rule-writing. In this work we compare these methods with each other and some baselines. We compare the effect of modality - text-based interaction against the spoken interaction with the embodied agent.

In the next section we will introduce our first case study system for unannotated corpus-based virtual human dialogue manager. In the next section we will elaborate more on the motivation for using corpus based methods for such systems. In section 4 we describe the chat-bot systems we have implemented. Section 5 presents the evaluation of the implemented systems and we conclude with discussion and future work.

2. SASO-ST

At USC's Institute for Creative Technologies, researchers have developed several prototype virtual human characters used for simulation training. SASO-ST [9] is one such environment, involving a prototype of a training environment for learning about negotiating with people from different cultures and with different beliefs and goals. In the first scenario, the trainee acts as an Army Captain negotiating with a simulated doctor. The trainee's goal is to convince the doctor to move his clinic to another location. The captain can offer help in moving the clinic and some other perks like medical supplies and equipments.

In order to investigate this domain, and build resources for the system, we collected a corpus of role-play dialogues and Wizard of Oz (WoZ) dialogues. Role-play dialogues feature more free-form human face to face interaction whereas the WoZ interactions are constrained by allowing the wizard playing the role of doctor to choose from a limited set of replies.

3. Motivation

Traditionally the development life cycle for a spoken dialogue system has involved a series of steps. Designers start by defining the domain of interaction and collecting human-human dialogue data to validate and refine it. This can followed by Wizard-of-Oz data collection where a human(wizard) takes part in conversation on behalf of the system. The main differ-

	Chat-bots	**Virtual Humans**	**Task Oriented**
Domain	unrestricted	somewhat restricted	restricted
Goal	be human-like	be human-like and complete task	complete the task efficiently
Understanding	shallow or no understanding of progression of dialogue needed	shallow understanding of dialogue progression needed	deep understanding of dialogue progression needed
Operating level	surface text	surface text or dialogue act	dialogue-act
Method	keyword-spotting, pattern matching, corpus based retrieval	information-state based or corpus based retrieval	Information-state based, form based
e.g.	Eliza, Alice	SASO-ST, Sgt Blackwell	Trains, Communicator

Figure 1: Various types of dialogue systems

ence between WoZ and roleplays is that the interaction with the wizard is more restricted e.g. Only pre-determined utterances/templates can be used as responses. All this dialogue data is transcribed and can be used for domain adaptation of acoustic and language models for speech recognition. For task oriented dialogue systems that operate on the dialogue-act level, the data is annotated for dialogue-acts and other semantic content. Although there have been some efforts in automatic dialogue-act tagging these are limited in applicability mostly due to the fact that the set of dialogue-acts heavily depends on the application domain. For rule-based systems like those employing information-state [10] based modelling, rules are written that operate on the input consisting of dialogue-act, semantic information and previous information-state. The process of annotating the collected dialogue data and authoring the rules for updating information-state serve as bottlenecks in the process of building dialogue systems.

In this paper, we report on initial efforts in solving this bottleneck by avoiding the higher abstraction level of dialogue-acts and remaining at a textual level. With a focus on virtual human dialogue systems, we explore the effectiveness of simple techniques in rapidly building dialogue models. Our goal is to build dialogue systems with minimal or no annotation. Our methods are inspired by Information Retrieval and work with the main assumption that a dialogue system can come up with a response by retrieving the utterance rather than constructing one from an abstract representation. Retrieval of pre-constructed utterances are also often used for constructing WoZ systems. This same retrieval strategy is often used in animation of virtual human bodies using motion capture rather than from procedural physics. Our methods work at the surface text level and retrieve the utterances from a corpus of un-annotated human-human dialogues.

4. Unsupervised dialogue models

We view the problem of dialogue modelling simply as predicting the most appropriate utterance given the context. We also assume that since the training data is human-human interaction, the most appropriate is also the most probable according to the training data. So the task is to find,

$$utt = argmax_i P(utt_i | context)$$

We examine several different dialogue modelling algorithms to find the best utterance.

4.1. random

In order to establish the lower baseline we implemented a classifier which returns a random utterance from the corpus. There were 435 utterances in the training set with doctor as the speaker. This bot does not capture any context.

4.2. nearest context

This bot tries to capture the local context. It implements a nearest neighbor classifier for predicting the most probable utterance. Here the context is approximated by the previous n turns $(n = 2)$. Following the vector-space model [11], the context is represented by tf-idf weighted vector of the words that occurred in previous n turns. The features used are stemmed unigrams augmented with speaker and distance in time in units of turns. The weighing scheme is modified to reflect the importance of immediately preceding turn.

Let W_i^j be the weight assigned for unigram w_i which appears j turns ago. Then W_i^j is given by,

$$W_i^j = TF(w_i) \times IDF(w_i) \times H(j) \tag{1}$$

$$TF(w_i) = 1 + \log(\#w_i) \tag{2a}$$

where $\#w_i$ is the number of times w_i appears in the utterance

$$IDF(w_i) = \log\left(\frac{N}{df_i}\right) \tag{2b}$$

where N is the total number of utterances
and df_i is the number of utterances containing w_i

$$H(j) = \exp\left(\frac{-j^2}{2}\right) \tag{2c}$$

Fig 2 shows an example of the feature vector used to represent the context. From the corpus we extract the pairs $(utt_i, context_i)$ where $context_i$ is the feature vector for the context corresponding to the utt_i. When its time to predict the utterance for the given $context$, we find $context_i$ such that distance between $context$ and $context_i$ is minimized. The utterance utt_i corresponding to the $context_i$ is then chosen as the reply. Fig 3 shows an example interaction.

4.3. segmented nearest context

Approximating the context by previous n turns is not ideal and suffers from problems that result due to ignoring long distance relationships. One such problem is the violation of presupposition as illustrated by last utterance in Fig 3. A presupposition is an assumption that must be made by dialogue participants for an utterance to make sense. e.g. when the Doctor says “ I don’t want to move the clinic, ” this makes sense only if the the dialogue participants both believe that it is relevant for the doctor

doctor	what do you want i have patients waiting for me
captain	I have orders to assist you in moving this clinic

<captain>_0_orders	6.24610677
<captain>_0_assist	4.16666522
<captain>_0_you	0.56935296
<captain>_0_moving	2.98801023
<captain>_0_this	1.7327793
<captain>_0_clinic	2.41746537
<doctor>_1_what	1.21161264
<doctor>_1_do	1.18616006
<doctor>_1_want	1.86086817
<doctor>_1_patients	1.02003814
<doctor>_1_waiting	2.45577118

Figure 2: feature vector representing the context of previous *n* turns. *(here n=2)*

captain	hello doctor
doctor	hello
captain	i am captain kirk nice to meet you
doctor	i'm uh doctor perez uh i'm very busy right now i i only have uh just a few minutes to uh
captain	i want to talk to you about an important issue
doctor	uh captain i i appreciate you have given us a lot of very good information to uh which roads are safe and where the landmines are and i need you+ uh i i cannot move this location though because of all these patients they're they're too critical right now i'm working on a on a on a young girl with amoebic dysentery and and she she requires my attention at all times i there is no way i i+ these people are in no shape to to+

Figure 3: example interaction for nearest context bot

to consider moving the clinic. And in this case it comes down to whether or not moving the clinic was previously suggested.

To minimise the violations of presupposition we keep track of which key concepts have been talked about in the dialogue. These concepts are manually identified along with the words that indicate their presence as in Fig 4. Tracking these concepts effectively segments the dialogue into various phases. Each context is now also augmented with the state (mentioned or unmentioned) of key concepts (or segment-signature). By just tracking which words have appeared in the dialogue we can identify the segment-signature. For choosing the best reply given the context we use similar mechanisms as for the nearest context bot, but restrict the set of utterances which are candidates for retrieval to only those that share the same segment-signature as the current context. Fig 5 shows an example interaction. This bot captures local as well as global context.

4.4. segmented random

This bot captures only global context. It keeps track of the key-concepts as mentioned earlier and randomly retrieves utterances from the set with matching segment-signature.

[move]	move , relocate
[supplies]	antibiotics, supplies medicines, plasma
[transportation]	vehicle, trucks transportation

Figure 4: list of key concepts along with the representative unigrams compiled for SASO-ST dialogues

captain	hello doctor
doctor	hello
captain	i am captain kirk nice to meet you
doctor	i'm uh doctor perez uh i'm very busy right now i i only have uh just a few minutes to uh
captain	i want to talk to you about an important issue
doctor	ok

Figure 5: example interaction for segmented - nearest context bot

5. Evaluation

In our previous work [8], we implemented text based versions of the four bots described here. We asked volunteers to engage in negotiation with the simulated doctor. Input and output modality was limited to text only and the turns were strictly alternated. We had six volunteers chatting with all the bots. The presentation order of the bots was balanced. We have 24 dialogues from these text-based interactions.

To understand the effect of speech recognition, we also collected dialogues from spoken interactions with an embodied character. The body, voice, and recognition and gesture production components from the SASO-ST doctor were used, with a dialogue manager based on random and segmented nearest neighbor conditions. Four volunteers talked to each version, with balanced presentation order. These spoken dialogues were later transcribed and the word error rate for the random version was 0.52, while the segmented nearest was 0.41.

We evaluated all the dialogues from both the previous study [8] and the current one. We also chose 4 human-human dialogues from our training corpus to establish the upper baseline. We had two evaluators judge the doctor's utterances for appropriateness on a scale of 1 to 5, with 1 being a totally non-sensical response and 5 a highly appropriate one. Evaluators used only the text transcriptions to make their judgements. We used the average of the two judgements as the final rating. Fig 6 presents the average appropriateness levels for different bots in different settings. Table 7 summarizes the results.

We performed Wilcoxon rank sum test to check whether the differences are statistically significant. Every other system was significantly better than text-random and speech-random. Human dialogues were significantly better than all the systems. Text based segmented nearest context was significantly better than segmented random but not better than nearest context. So even though segmentation helped it was not significant. At the same time speech did have a significant effect in lowering the perceived appropriateness.

Fig 8 shows the scatter plot of average ratings for the bots as judged by our two judges. The linearity of the plot suggests high inter-rater agreement, even though one rater tended to give much higher scores across the board. As a measure of inter-rater agreement we calculated the pearson's correlation coefficient

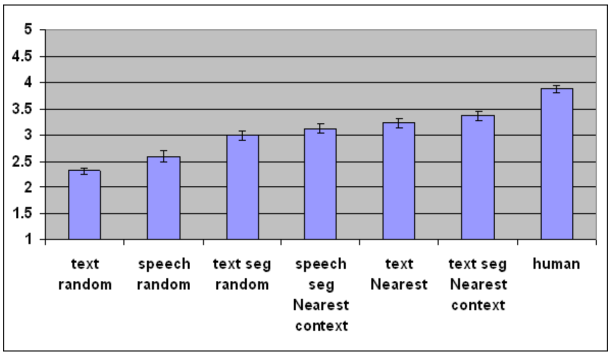

Figure 6: Average appropriateness levels for different bot types

Bot	# of utterances	avg judge1	avg judge2	avg both	stderr
text random	141	2.75	1.86	2.30	0.070
speech random	88	3.14	2.06	2.60	0.102
text seg random	96	3.52	2.47	2.99	0.088
speech seg nearest context	99	3.47	2.78	3.13	0.096
text nearest	103	3.85	2.60	3.23	0.092
text seg nearest context	113	3.96	2.78	3.37	0.091
human	91	4.38	3.36	3.87	0.074

Figure 7: Results for various types of chat-bots

for average appropriateness levels. It is quite high at 0.94.

6. Conclusion and Future Work

In the preliminary evaluation and the subjective feedback from the users it appears that segmented nearest context system performs surprisingly well. Retrieving the utterances rather than generating them also adds more richness and naturalness and to the replies for virtual human, making it more believable.

Currently the segmentation is based on manual identification of concepts. In future, we will try to automatically identify the key concepts used to segment the dialogue, as well as looking at what information would both improve dialogue quality and be able to be extracted automatically or authored with little effort. We will also investigate how these methods can be applied to tasks which have a more deeper structure. We are also interested in automatic methods to evaluate these types of systems.

7. Acknowledgments

We would like to thank the members of the SASO-ST project for making the virtual human components and corpus available for this work. The 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.

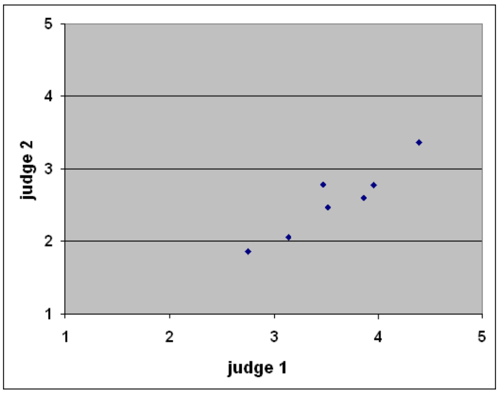

Figure 8: Scatter plot of average appropriateness levels for different bot types as judged by two judges

8. References

[1] J. Weizenbaum, “Eliza–a computer program for the study of natural language communication between man and machine,” *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, January 1966.

[2] R. Wallace, *Be Your Own Botmaster, 2nd Edition*. ALICE A. I. Foundation, 2003.

[3] E. Seneff, L. Hirschman, and V. Zue, “Interactive problem solving and dialogue in the atis domain,” pp. 354–359, February 1991.

[4] J. F. Allen, “The trains project,” *Journal of Experimental and Theoretical AI*, 1995.

[5] J. Chu-Carroll and B. Carpenter, “Vector-based natural language call routing,” *Journal of Computational Linguistics*, vol. 25, no. 30, pp. 361–388, 1999.

[6] A. Leuski, R. Patel, D. Traum, and B. Kennedy, “Building effective question answering characters,” in *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, 2006.

[7] B. A. Shawar and E. Atwell, “Using dialogue to retrain a chatbot system,” in *In proceedings of corpus linguistics*, 2003.

[8] S. Gandhe and D. Traum, “First steps towards dialogue modeling from an un-annotated human-human corpus,” in *5th Workshop on knowledge and reasoning in practical dialogue systems*, Hyderabad, India, 2007.

[9] D. Traum, W. Swartout, J. Gratch, and S. Marsella, “Virtual humans for non-team interaction training,” in *AAMAS-05 Workshop on Creating Bonds with Humanoids*, July 2005.

[10] D. Traum and S. Larsson, “The information state approach to dialogue management,” in *Current and New Directions in Discourse and Dialogue*, J. van Kuppevelt and R. Smith, Eds. Kluwer, 2003.

[11] C. Manning and H. Schutze, *Foundations of Statical Natural Language Processing*. MIT Press. Cambridge, MA, 1999, ch. 15.