Thesis Proposal:
An Investigation on The Effectiveness of Employing Topic Modeling Techniques to Provide Topic Awareness For Conversational Agents

Omid Moradiannasab

Saarland University

66123 Saarbrcken, Germany

`omidm@coli.uni-saarland.de`

Abstract

The idea behind this proposal is to investigate the possibility of utilizing NLP tools, statistical topic modeling techniques and freely available online resources to propose a system able to provide dialogue contribution suggestions which are relevant to the context, yet out of the main activity of the dialogue (i.e. *off-activity talk*). The aim is to evaluate the effects of a tool that automatically suggests *off-activity talk*s in form of some sentences relevant to the dialogue context. The evaluation is to be done over two test-sets of open domain and closed-domain in a conversational quiz-like setting. The outcome of this work will be a satisfactory point of entry to investigate the hypothesis that adding *automatically generated off-activity talks* feature to a conversational agent can lead to building up engagement of the dialogue partner(s).

1 Introduction

Conversational agents (e.g. virtual characters, chat-bots, etc) are software programs that interact with users employing natural language processing capabilities. The ability to interact with user in natural language enables such automated agents to make task-oriented dialogues with the aim to provide services to human in different fields such as education, entertainment, help desks, etc. *Traditional conversational systems* are designed for specific purposes, such as a banking service or navigating through a website. Dialogue in such systems is limited to *task-bound talks* (also called *activity talks*). These talks have a specific purpose and follow a particular structure. *Traditional conversational systems* are developed to attempt to engage the user in a natural, robust conversation in well-defined domains. However, empirical investigations reveal that the effect of these systems on engagement of the users with the system and their perception of the agent's intelligence is debatable (Dehn and Van Mulken, 2000). *Relational agents*, on the other hand, are defined in literature as "computational artifacts designed to establish and maintain long-term social-emotional relationships with their users" (Bickmore and Picard, 2005).

In order to achieve relational agents, certain amount of user's trust and engagement is required. Various conversational strategies are employed in *relational agents* that comprise models of social dialogues with the aim of raising user's trust. The conversational strategy targeted in this work is *off-activity talk*. An *off-activity talk* is a verbal reaction which is required to be contextually relevant to the content of the previous interaction and preferably includes a provision of some added-information. This verbal reaction can consist of one or more sentences and is extracted from the freely available online resources. The output of this work can be employed as a generator of relevant utterances embedded in an artificial agent (i.e. a virtual character or a robot). The sub-tasks addressed in this work are to analyze the dialogue context, to detect the topic, and to provide a ranked list of appropriate candidate utterances to be employed by the dialogue manager in a conversation system.

With the presented idea as a starting point, advantages or disadvantages of including automatically retrieved OATs in conversation systems can be studied. It is encouraging to investigate the effect of embedding such system on engagement of the users with a conversation system. It is also desirable to investigate whether automatically retrieved OATs can promote users' trust in knowl-

edgeability of the agent and their perception of the agent's intelligence.

2 Background

2.1 Small Talk

As an instance of a conversational strategy employment, *small talk* (also *social talk*) is discussed in literature. It is introduced as a kind of talk which executes conversational strategies. While interleaved between *task-bound talks*, *social talk* indirectly builds trust through the natural progression of a conversation. Bickmore and Cassell (2001) define *small talk* as "any talk in which interpersonal goals are emphasized and task goals are either non-existent or de-emphasized". According to their work, "One strategy for effecting changes to the familiarity dimension of the relationship model is for the speaker to disclose information about him/herself and induce the listener to do the same". Klüwer (2015) defines it as a talk "often perceived as unsophisticated chit-chat in which content exchange is irrelevant and negligible". Following this definition, *small talk* (e.g. about the weather, events and objects around, stories about the environment or the virtual character itself) represents the opposite of *task-bound talk* which aids the execution of a particular task. Thus, the range of topics and contents is definitely much more unrestricted in *small talk* than in *task-bound talk*. *Small talk* is useful to develop the conversation and to avoid pauses. It can be used to ease the situation and to make a user feel more comfortable in a conversation with an agent (Cassell and Bickmore, 2000). It is also introduced as a way of assuring certain amount of *closeness* to the user (e.g. before asking personal questions) (Bickmore and Cassell, 2001). *Small talk* is also helpful to avoid repetitiveness of conversations which is counted in literature as a negative impact factor to the users' motivation in interaction with the agent.

2.2 Off-Activity Talk (OAT)

Similar to *small talk*, *off-activity talk* (also called *non-activity talk*) is another technique to employ conversational strategies. Both *small talk* and *off-activity talk* enrich the task-oriented dialogue via opening the structure of the conversation. However, OAT can be differentiated from *small talk* by the topic and the purpose of the talk. OAT has a specific purpose (e.g. knowledge exchange) and is about a specific topic (see Table 1), while a *small talk* is an independent talk without any functional topic (e.g. talking about the weather). Several studies have found that the purpose of *small talk* is not to negotiate knowledge but to aid in management of social situation. On the other hand, *off-activity talk* is to disclose some information relevant to the dialogue context. Therefore, whenever no divergence from the subject matter of the *task-bound talk* is required, OAT is preferred to *small talk*. Even though OAT is a diversion from the structure of the task-dialogue, it maintains the dialogue topic.

Table 1: Some samples of a suitable OAT

Sub-dialogue(A+B) And Follow-up OAT(C)
A: What is the capital of Chechnya?   B: Grozny   C: Not long ago, Grozny, the capital of Chechnya, was called the most devastated city on Earth.
A: What is the capital of Chechnya?   B: Grozny   C: Chechnya Republic is a federal subject of the Russian Federation, part of North Caucasian District.
A: When was the Berlin Wall built?   B: in 1961   C: Originally a barbed wire fence, the first concrete sections were built in 1965.
A: When was Elvis Presley's first record recorded?   B: on July 5, 1954   C: On July 5 1954, Elvis Presley changed music world forever.
A: How many portions of fruit and vegetables should we try to eat?   B: At least five a day   C: Healthy diet means 10 portions of fruit and vegetables per day, not five.

OAT was first defined by Kruijff-Korbayova et al. (2014) in resemblance to *small talk*. In their work, the purpose of *activity talk* is knowledge exchange or knowledge probing while *off-activity talk* is used to break out of the fixed structure of task-bound dialogues. They use OAT in form of prerecorded questions around a set of predefined topics (e.g. hobbies, diabetes, eating habits, friends, diary, etc) to encourage the user to talk about those topics in order to elicit information from them. In this study, the focus is on *off-*

activity talk with a similar definition but a different employment. An OAT in the current work is a follow-up added information relevant to the context of the previous interaction with the aim of encouraging users' engagement and possibly promoting their trust in knowledgeability and intelligence of the agent, thus no deliberate direct information elicitation is targeted in this work.

3   Use Case Scenario

The idea of context determination and relevant off-activity talk suggestion for dialogue contribution, in a broad sense, can be used in any task-oriented dialogue setting. However, due to its predominantly verbal character and naturally constrained interaction structure, a conversational quiz-game setting is chosen as a good test bed for the current work. Some examples of dialogues in this setting, in addition to some sample OATs to be uttered by the agent right after this dialogue, are presented in Table 1. In this scenario, the agent asks the user a multiple-choice question from an open domain (A). After the user selects one of the choices (which can be correct or not) (B), the agent should give a verbal reaction (*off-activity talk*) (C). The follow-up needs to be related to the content of the previous interaction. It should be a piece of information on the main subject matter extracted from the available online resources and possibly, but not necessarily, confirm or give the correct answer. It can include a provision of some added information as a follow-up to the previous content.

4   Research Objectives

The goal of this project is to evaluate a tool which takes advantage of the Web as an online resource to seek for relevant sentence(s) to a given dialogue context. A quiz-game setting is used as the test bed for this project. The main benefit of such a tool will be to break out of the fixed structure of task-bound dialogues between an artificial agent and a human user, ideally leading to an increment in users' engagement. *Off-activity talk*s suggested by this tool can be employed in an online or offline mode. In the offline mode, OATs will be used to create (or enrich) a handcrafted knowledge base for a virtual character. However, offline enrichment might not be feasible in open domain scenarios. That is why the author puts the ultimate aim of this work to provide an open-domain solution for real-time OAT suggestion.

In the process of this study, following research questions are to be answered:

1. What are the features of a suitable off-activity talk?

2. Which of the proposed approaches is more effective in providing a high quality sentence selection?

3. How far can one reach by employing topic modeling techniques in the proposed tool?

4. To what extent using online resources can help with breaking out of the fixed structure of the task-bound dialogue?

5. Does providing OATs via sentence selection increase user satisfaction?

6. Does providing OATs via sentence selection boost user's trust in knowledgeability and intelligence of the agent?

7. Does providing OATs via sentence selection increase semantic cohesion of the dialogue?

5   Method

At large, the procedure is as follows. The first step is to identify topic-related *focus terms* of a dialogue. Using these *focus terms*, the second step is to determine the major topic to be followed up in the conversation. The topic labels after this step do not necessarily need to match the terms that occurred within the previous dialogue. In closed-domain solutions, this can be done by mapping the *focus terms* onto a category taxonomy. In such a case, the taxonomy is utilized as a reference point to derive topic labels for a follow-up conversation. On the other hand, for an open-domain problem the task is not as straightforward. The reason to that is the lack of predefined taxonomy. Topic modeling techniques are employed to determine the major topic of the conversation (see 5.1). Identified topic categories are subsequently combined with information retrieval and different linguistic filtering methods to improve candidate content retrieval. In the end, the retrieved content will be ranked based on their relevance to the context and properness for dialogue contribution.

5.1   Topic Modeling

Topic Modeling, as a branch of text mining, is the process of identifying patterns in the text in order

to classify words into groups called "topics". A topic is defined as a probability distribution over the terms in the vocabulary. In this process, topics are assigned to documents and terms are assigned to topics each with specific probability distributions. There are different methods to achieve this goal: LDA, HDP, NNMF, etc.

6 Evaluation Strategy

The main goal of the experimental evaluation in this study is to assess the potential of the proposed system in contributing to a dialogue with two distinct partners: a questioner and an answerer. In this quiz-like setting, the output of the system is supposed to be an utterance by a conversational agent which follows the dialogue by elaborating on that and by providing relevant information in an appropriate way for a real-world dialogue contribution.

A dialogue system can be evaluated in various styles. The evaluation approach can be either subjective or objective (Walker et al., 1997). Evaluation metrics can be derived from questionnaires or log files (Paek, 2001). The scale of the metrics can vary from the utterance level to the whole dialogue (Danieli and Gerbino, 1995) (Kamm et al., 1999). The dialogue system can be treated as a "black box" or as a "transparent box" (Hone and Graham, 2000). This variety of styles beside lack of any agreed-upon standards in the research community and incompatibility of evaluation methods make evaluation of dialogue systems a challenge.

6.1 Embedded or Stand-alone Strategy?

An ideal strategy to test a tool, which automatically suggests *off-activity talks* for dialogue contribution, is embedding the developed component in a test-bed application and assessing the change in the usability of that application by carrying out a subjective evaluation of user satisfaction. However, it is not always easy to get access to such platforms and to perform the required embedding scheme. An alternative is to assess this application in a stand-alone scheme. There are some advantages to assessing with a stand-alone scheme instead of an embedded one. As an example, by employing a stand-alone scheme, one can avoid problems which might occur during the interaction with the embedding system. Such problems can influence the usability results while we have no control over them. Therefore, the stand-alone

strategy, in this case, brings the benefit to focus specifically on the quality of the OATs. The author recommends to observe the inputs and outputs of the system and while treating the system as a "black box", to measure the plausibility of the proposed approach. This also means that the measurement scale in the evaluation scheme is at utterance level, in contrast to dialogue level.

6.2 Subjective or Objective Evaluation?

As stated earlier, evaluation of a dialogue system can be done with an objective or subjective approach. In case of objective evaluation, metrics like resources used (e.g. time, turns, user attention, etc) or the number of errors the system makes or inappropriate utterances made by the system can be mentioned. In some cases, a number of specified definitions of task success is used as an objective metric. However, it is not always easy to define task success in an objective way. The subjective measurement of the acceptance of an application or technology belongs to the group of usability evaluation. Usability evaluation focuses on users and their needs. Through usability evaluation, we want to figure out if a system can be used for the specific purpose from the user's point of view, and if it allows the users to achieve their goals, meeting their expectations. The most important criterion for measuring usability is user's satisfaction. The author proposes to assess the usefulness of the suggested OATs from the tool by measuring user satisfaction with regard to a specific factor. That means, the evaluation strategy falls into the category of subjective evaluation.

Information about user satisfaction is usually gathered through interviews and questionnaires in the end of a session of interaction with a dialogue system. In principle, the goal of this work is not to establish a dialogue with the human, but to create a component to be integrated in a dialogue system and provide suggestions to the dialogue manager of such a system. That is why, it is not possible to evaluate user's satisfaction in the end of a dialogue session. Alternatively, we can ask participants to qualify the suggested OATs separately and regardless of any potential preceding or following dialogue. In other words, the experiment is not designed as a dialogue; instead, pairs of inputs and outputs of the system are presented to participants and they are asked to define which output is more satisfactory regarding some well-described

aspects.

In contrast to objective evaluation techniques which are fairly well-established, subjective measurements are not as structured and straightforward. A difficulty which arises here is that dialogue systems and their users sometimes have inconsistent attitudes toward a dialogue. As a consequence, extra care needs to be taken to make sure that the participants of the experiments have a correct sense of what is needed to measure in a qualification process. Different factors must be defined with regard to which the quality of an OAT is to be measured. So, It is needed to provide accurate instructions to make sure the participant has comprehended the differences and is able to distinguish between these factors. The author also proposes to use some test questions as a means to make sure that the participants have read and clearly understood the instructions.

6.3 Evaluation Factors

In order to judge the usefulness of an OAT, it is needed to define the aspects of evaluation. The type of application determines the aspects that are important for a usability evaluation. To decide about the main aspect of the evaluation, the author largely relies upon SASSI methodology (Hone and Graham, 2000). SASSI is a methodology for evaluating spoken dialogue systems. They state that the previously reported subjective techniques are unproven. Also, It is argued that their content and structure are, for the most part, arbitrary and the items chosen for a questionnaire or rating scales are based neither on theory nor on well-conducted empirical research. The reasons for choosing a particular structure in the previous studies (e.g. questions, statements or numerical scales) and sub-structure (presentation, number of points on a scale, etc.) are not reported. Therefore, they use factor analysis to determine the main components of users' attitude and also define suitable rating scales for each of these components. Resultant factors after labeling are:

1. response accuracy

2. likability

3. cognitive demand

4. annoyance

5. habitability

6. speed

The first factor (response accuracy) does not fully match this application. The reason is that accuracy or the number of errors refer to the users' expectation of what the system is supposed to do in response to the input utterance. So, a specific goal like acting appropriately should be defined which is not necessarily the case with a free off-activity talk. The third group (Cognitive Demand) refers to the perceived amount of effort needed to interact with a dialogue system and the feelings resulting from this effort. Since in the current case the participant is not actually talking to the system but rating some contributions in a conversation, therefore it is not suitable to ask their opinion about how it will feel to talk to it or how much effort will be needed to say something easily comprehensible to the system. The fifth group (habitability) also refers to whether the user knows what to say to the system at each turn which is not useful as reasoned for the third group. The sixth factor (speed) is also not proper because the participants do not interact in real-time to the system and only rates some logged interactions. So, they do not know how fast the system works.

Among the factors mentioned in SASSI, likability and annoyance are applicable for the current experiment. In addition, questions of some other features are also important to us (e.g. naturalness, repetitiveness, etc). Altogether, the author proposes the items in Table 2. For each appropriateness judgment, this list of statements is disclosed to the participants and they are asked to check those which are true in their subjective opinion.

Table 2: Appropriateness judgment statements

Judgment Statement
This is a natural response.
Talking to a system with such a response is boring.
I would enjoy talking to a system which responds like this.
This talk is repetitive.
Such an utterance by a robot is frustrating.
The system knows how to speak to the human.
The follow-up provides relevant added-information.

84

References

Timothy Bickmore and Justine Cassell. 2001. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 396–403. ACM.

Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327.

Justine Cassell and Timothy Bickmore. 2000. External manifestations of trustworthiness in the interface. *Communications of the ACM*, 43(12):50–56.

Morena Danieli and Elisabetta Gerbino. 1995. Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI spring symposium on Empirical Methods in Discourse Interpretation and Generation*, volume 16, pages 34–39.

Doris M Dehn and Susanne Van Mulken. 2000. The impact of animated interface agents: a review of empirical research. *International journal of human-computer studies*, 52(1):1–22.

Kate S Hone and Robert Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (sassi). *Natural Language Engineering*, 6(3&4):287–303.

Candace Kamm, Marilyn A Walker, and Diane Litman. 1999. Evaluating spoken language systems. In *Proc. of AVIOS*. Citeseer.

Tina Klüwer. 2015. Social talk capabilities for dialogue systems.

Ivana Kruijff-Korbayova, Elettra Oleari, Ilaria Baroni, Bernd Kiefer, Mattia Coti Zelati, Clara Pozzi, and Alberto Sanna. 2014. Effects of off-activity talk in human-robot interaction with diabetic children. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, pages 649–654. IEEE.

Tim Paek. 2001. Empirical methods for evaluating dialog systems. In *Proceedings of the workshop on Evaluation for Language and Dialogue Systems-Volume 9*, page 2. Association for Computational Linguistics.

Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280. Association for Computational Linguistics.