Handling out of domains topics by a Conversational Character

Corradini, Andrea; Mehta, Manish

Published in: Proceedings of the 3rd ACM International Conference on Digital Interactive Media in Entertainment and Art

Publication date: 2008

Document version: Final published version


Go to publication entry in University of Southern Denmark's Research Portal

Terms of use
This work is brought to you by the University of Southern Denmark. Unless otherwise specified it has been shared according to the terms for self-archiving.

If no other license is stated, these terms apply:

• You may download this work for personal use only.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying this open access version

If you believe that this document breaches copyright please contact us providing details and we will investigate your claim. Please direct all enquiries to puresupport@bib.sdu.dk
ABSTRACT
Unlike traditional task-oriented dialog systems, one of the goals of conversational applications is to keep the user engaged and interested for as long as possible. This paper describes a technique that we implemented for dealing with the problem of out-of-domain topics and successfully integrated in a fully functional conversational system. In our approach, we exploit a simple and freely available coarse-grained ontology, namely Google’s directory structure, in order to automatically categorize unknown words and complete partial user utterances before further semantic processing. We also resort to available web-based question and answering (QA) systems to generate answers to user’s questions that are outside the domains covered by the application. Further, we make a first attempt at categorizing the retrieved information to generate appropriate non-verbal behaviors synced up with spoken utterances. The evaluation of the complete system shows promising results for the future and suggests that we are on the right track in terms of our approach for a better handling of out-of-domain topics.

Categories and Subject Descriptors
H.5 [Information Interfaces and Presentation (e.g., HCI)]: User Interfaces—natural language, Evaluation; K.3 [Computers and Education]: General

General Terms
Design, Human Factors, Experimentation

Keywords
Out-of-domain topics, multimodal conversational systems, QA systems, user interfaces.

1. INTRODUCTION
In recent years, a growing research community [7, 13, 21] has started working on development of conversational agent based interfaces which exhibit human-like behavior and appearance. These interfaces, termed embodied conversational agents (ECAs) [6] aim to use and realize cues inherently peculiar to human-human communication, such as sense of presence, mixed initiative and non-verbal behaviors to hold up their end of the dialog with the user. Most of the previous work on ECAs has centered on task oriented conversational systems helping the user accomplish a specific well-defined work in an effective manner and the discussion topics have been restricted to the task at hand. Typically out-of-domain input has been handled by either ignoring it or by having the system express its inability to properly operate on it [20]. These strategies are valuable in a task oriented system, where the goal is to keep the conversation focused on the topic of interest whereby out-of-domain input is irrelevant to and not essential for achieving the target objective. However, as we start developing ECAs with the clear goal in mind to provide the user with a rich social experience and act as a conversational partner rather than a mere computational task solver, effective handling of out-of-domain input becomes an important issue. We believe that simplistic approaches could hamper the overall player experience rather than enhancing it and would thus contradict with our original ideas and intentions.

Natural language processing and understanding for domain oriented conversation is a notoriously difficult problem; building a conversational system that can effectively handle out-of-domain input and provides a range of discussion topics is therefore even a harder issue. Such a complex task requires common sense reasoning and knowledge which human beings make use of in their everyday dealings with the world. Moreover, conversation on the different out-of-domain topics requires significant content creation skills, additional time to update the knowledge base and memory space. Ideally, we want an approach that takes into account the authorial burden involved in creating the content, uses an understanding approach that is able to deal with these different topics and simultaneously doesn’t require a lot of effort. Our research effort is aimed toward the development of approaches that achieve these objectives in the context of a real time conversational system that allows for fun and experientially rich social interaction between kids and an embodied agent through spontaneous speech and 2D gesture. This paper presents an approach to address general purpose topics such as movies, games, current news, food, famous places and personalities by both using web resources, notably Google’s
ontological resources to enhance the understanding capabilities of the system, and utilizing existing question-answering (QA) systems and resources freely available on web sites to address these topics.

The rest of the paper is organized as follows. We discuss the existing approaches on handling out-of-domain input in Section 2. We present a general overview of the system and understanding approach in Section 3. We propose a methodology that utilizes automatic categorization available through the Google’s directory structure to understand names of and information about topics that come up in everyday conversations, but are not included in the system knowledge base, to enhance the existing understanding approach in Section 4. We then discuss the conversational mover module that is used to detect the next prospective conversation move in Section 5. In Section 6, we present our method of using open domain QA systems and information available on the web to address multiple general-purpose topics. We present an evaluation of our approach in Section 7. Eventually, we provide a discussion of our current work in Section 8 and conclude in Section 9.

2. EXISTING APPROACHES

Current work on developing programs that can simulate typed conversation mostly rely on a template based approach to generate answers to multiple general purpose topics [15]. The interaction philosophy underlying these systems, also referred to as chatterbots, is very simple and effective at the same time. They maintain full control of the interactive session by keeping the conversational flow on a specific, well defined (from the system’s perspective) track and leave room for as less opportunities as possible for the human interlocutor to take the initiative. In doing so, they adopt a strategy that avoids in advance situations where the user could ask questions about or require responses related to an unconstrained range of utterances that will reveal the actual limitations of the system. Other approaches have used similar template based approach to address out-of-domain topics [19] and engage in small talk [4]. The range of discussion topics is still limited since it is dependent on the amount of templates that can be created off-line. Moreover, creating these set of templates requires hand crafted answers to all the possible imaginable discussion topics. In our approach, we want to reduce the authorial burden of content creation for different general purpose discussion topics.

In [2] an approach where web content is used to have a conversation on different general purpose topics is presented. Our approach differs from it because our system addresses general purpose topics by combining normal domain oriented conversation with the outcome of question answering systems and content posted on specific web sites. Another approach to handle out-of-domain input through a set of answers that explicitly state that the character either doesn’t know or doesn’t want to reveal the answer is proposed in [20]. Such an approach is in general better than saying something completely absurd, however this strategy is more suitable for training simulations where the goal of the system is to keep the conversation on track so as to achieve the training goal. For our domain where the goal of our agent is to provide an appropriate reply along with a rich social experience to the user, these strategies will not work.

Facade [14], an interactive drama domain, uses various deflection strategies to bring back the discussion onto the main conversation as well as limit the depth in which players can drill down on any one topic. These strategies present an interesting solution to avoid out-of-domain input for a story based domain. An ongoing study provides the user with enough narrative cues to integrate the deflection output used by characters into the ongoing narrative flow [17]. Differently from that work, in our approach, we want to address the general purpose topics apart from the domain topics rather than deflecting them to bring the conversation back onto the domain topics.

3. SYSTEM PRIMER

3.1 Background

An interactive computer game where a player can interact with an embodied character in a 3D world, using spoken conversation as well as 2D pen gesture is the scenario of our choice to merge e-learning and entertainment. The system has been developed and deployed in an existing game scenario [8] where players can communicate with a computer generated representation of the Danish author and fairy tales writer Hans Christian Andersen (HCA), famous for his fairy tales, which were not written merely for children but for adults as well. In this application domain, the goal is to have players learn about the writer’s life, historical period and fairy tales and have fun at the same time hence making them play again and in turn learn even more.

There is no visible user avatar as the user perceives the world around him in a first-person perspective. She can explore HCA’s study and talk to him, in any order, about any topic within HCA’s knowledge domains, using spontaneous speech and mixed-initiative dialog. The user can change the camera view, refer to and talk about objects in the study, and also point at or gesture to them. Typical input gestures are markers like, e.g., lines, points, circles, etc. entered at will via a mouse compatible input device or using a touch-sensitive screen. HCA’s domains of discourse are: HCA’s fairy tales, his life, his physical presence in his study, the user, HCA’s role as gate-keeper for access to the fairy tale world, and the meta-domain of solving problems of meta-communication during speech/gesture conversation [8]. Apart from engaging in conversation on his domain of expertise, the user can discuss everyday topics like movies, games, famous personalities and others using a typed interface. In order to facilitate our later discussion on understanding general purpose topics, we first present the natural language understanding (NLU) approach used in our system.

3.2 NLU approach

The NLU module (see Figure 1) consists of four main components: a key phrase spotter, a semantic analyzer, a concept finder, and a domain spotter. Any user utterance from the speech recognizer is forwarded to the NLU where a key phrase spotter detects multi-word expressions from a stored set of words labeled with semantic and syntactic tags. This first stage of processing usually is helpful to adjust minor errors due to misrecognized utterances by the Speech recognizer. Key phrases that are domain-related are extracted, and a wider acceptance of utterances is achieved. The processed utterance is sent on to the semantic analyzer. Here,
Table 1: Example of processing inside the NLU.

<table>
<thead>
<tr>
<th>NLU Submodule Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech Recognizer</td>
</tr>
<tr>
<td>Keyphrase Spotter</td>
</tr>
<tr>
<td>Semantic Analyzer</td>
</tr>
<tr>
<td>Concept Finder</td>
</tr>
</tbody>
</table>

Table 1: Example of processing inside the NLU.

Dates, age, and numerals in the user utterance are detected while both the syntactic and semantic categories for single words are retrieved from a lexicon. Relaying upon these semantic and syntactic categories, grammar rules are then applied to the utterance to help in performing word sense disambiguation and to create a sequence of semantic and syntactic categories. The rule engine rewrites certain semantic/syntactic categories (or category sequences) in terms of other semantic/syntactic categories (or category sequences). This higher-level representation of the input is then fed into a set of finite state automata, each associated to a predefined semantic equivalent according to data used to train the automata. Anytime a sequence is able to traverse a given automaton, its associated semantic equivalent is the semantic representation corresponding to the input utterance. At the same time, the NLU calculates a representation of the user utterance in terms of dialog acts. At the next stage, the concept finder relates the representation of the user utterance, in terms of semantic categories, to the domain level ontological representation. Once semantic categories are mapped onto domain level concepts and properties, the relevant domain of the user utterance is extracted. The domain helps in providing a categorization of the character’s knowledge set. The final output in form of concept(s)/subconcept(s) pairs, property, dialog act and domain is sent to the dialog module through input fusion with the gesture modality. Table 1 shows an example of processing inside the NLU.

Generic rules are defined inside the semantic analyzer for detecting dialog acts. These dialog acts provide a representation of user intent like types of question asked (e.g., asking about a particular place or a particular reason), opinion statements (like positive, negative or generic comments), greetings (opening, closing) and repairs (clarification, corrections, repeats) [16]. These dialog acts are reused across different domains of conversation. Moreover, generic rules are used to detect the domain independent properties (e.g., dislike, like, praise, read, write etc). This sharing across different domains of conversation allows for faster understanding on new domains. As explained in the next section, these dialog acts are combined with domain independent properties and Google’s directory categorization for unknown concepts to achieve a semi-automatic categorization for general purpose topics.

4. GENERAL PURPOSE TOPICS THROUGH GOOGLE’S DIRECTORY

Web directories represent large databases of hand-selected and human-reviewed sites arranged into a hierarchy of topic categories. Search engines utilize these directories to find high-quality, hand-selected sites to add to their database. Users that are searching for a variety of sites on the same topic also find directories helpful by being able to search in only the category that interests them. The web directory of Google contains, among other things, classification information about names of movies, games, and famous personalities. Making entries for these domains manually in the lexicon would be a labor and time intensive effort. Apart from that, these open ended domains evolve over a period of time and need periodic updates. Thus, using Google’s categorization provides an automatic classification method for terms related to these domains. In our architecture, the NLU categorizes the word(s) without a lexical entry and those that are not detected by the keyphrase spotter, into an unknown category. The longest unknown sequence of words is combined into a single phrase. These words are sent to the web agent, which uses Google’s directory structure to find out whether the unknown words refer to a name of a movie, game, or a famous personality and the corresponding category is returned to the NLU. To illustrate the processing let us assume the user asked “do you like quake”. In this case, the NLU marks the word quake as an unknown category that, as such, needs further resolution. The temporary output of the NLU is thus a yes/no-question as dialog act, a property of the kind like and an unknown category. The unknown category is resolved by the web agent into the category game using Google’s directory engine (Fig. 2). Using this newly gathered information, the NLU is capable to pass on to the dialog module a complete output which now
consists of yes/no-question dialog act, a property of kind like, a concept game and a sub concept quake. Based on this information, the dialog module searches for an appropriate conversational move in response to the original sentence. The classification provided by Google along with the properties shared across domains and the dialog acts provides a method to build an automated representation consistent with the current output representation provided by the understanding module.

5. CONVERSATIONAL MOVER

At the next stage, inside the dialog module, the output representation from the NLU is used to reason about the next conversational move of the character. This stage of processing is performed inside a module called the conversational mover. For each conversational move of the character, rules are defined using the concept(s)/sub concept(s), property(s)/property type and dialog act/dialog act type pairs delivered by the NLU. This provides a systematic way to connect the user intention to the characters output move. Table 2 shows examples of rules inside the conversational mover. Anytime HCA has to produce a response or initiate a new conversational turn on the domain topics, the dialog module selects a contextually appropriate output in accordance with the conversational move produced by this module, the conversational history and the emotional state [8]. For output moves related to general purpose topics the dialog module retrieves a reply from different web sites and question answering systems. See next section for the details.

6. OUTPUT RETRIEVAL USING THE WEB AND QA SYSTEMS

Question-answering systems work fine in restricted domains [18]. In more complex situations, research in that field has reached a stage where they can even take an open domain natural language query and provide a simple answer [1, 10].

On the one hand, the creation of canned templates and natural language responses for open ended domains is an enormous and nearly impossible engineering exercise. On the other hand, the web stores a huge amount of information in HTML pages. Hence, one can expect to find the answer to a question already posted on the net. One only needs to find the right way to retrieve it. In our approach, we make use of existing QA systems that process the data available on web sites. In this way, we do not need to take care of the way information is gathered but only of the way we further process it once it is passed on to our system.

When the conversational mover classifies the NLU output representation into a conversational move whose output is to be retrieved from the web, the request is sent to the web agent. The web agent, depending upon the type of move, finds a quick and concise output using three freely available open-domain QA systems: AnswerBus[23], Start[11], and AskJeeves1 or the web pages at game2 and movies3 web sites (Fig. 1). The web agent employs a set of heuristics, such as removing output with certain stop words, to pick one single reply. Once a sentence is selected, we remove control/graphical characters to get a plain string that can be played by the Text-to-Speech (TTS) through the system's response generator component. This verbal output still doesn't have any non-verbal behavior assigned to it. We process this string to detect the semantic categories and dialog acts present in it. For carrying out this task we utilize the same key phrases, lexicon and grammar resources used by the NLU. In this case however, pre-stored mappings from semantic categories and dialog acts onto a set of non-verbal behaviors help attaching suitable non-verbal behaviors to the input segments that originally could not be processed.

For instance, when the user asks "do you know Pete Sampras" the NLU resolves the unknown words "Pete Sampras" through the Google's directory structure. The ontological representation of the user utterance is then sent to the conversational mover which classifies it into a move of the kind famous_personality_knowledge. This in turn, is then sent to the web agent which resorts to the services of the QA agents to retrieve information about Pete Sampras. After some post processing, the reply string, being in this case "Pete Sampras is an American tennis champion, ranked as the number one player in the world for several years in the

---

1www.ask.com
2www.game-revolution.com
3www.rottentomatoes.com
1990s” is sent to the system’s response generator. This output is further analyzed to find the semantic categories and dialog acts. During the analysis, the key-phrase “number one” is categorized into a semantic category “praise”, which in turn maps onto the non-verbal behavior “thumbs up”. A mapping from semantic categories onto non-verbal behaviors is defined at design time. All this results in the system’s response generator creating a graphical animation “thumbs up” while simultaneously playing back the corresponding verbal segment with the TTS. This methodology helps us assign some non verbal behaviors to the open ended verbal output produced through QA systems and web sites.

A current limitation of such an approach is that it is currently not possible to assign all the appropriate behaviors for a verbal output as coming up with a scheme that detects all the possible semantic categories and mapping the semantic categories onto non-verbal behaviors for open ended text is an enormous exercise.

7. EVALUATION

In order to measure the effectiveness of our approach, two independent evaluators, external to the project, conducted an evaluation study. They analyzed the set of out-of-domain inputs from two earlier studies [3, 8] that we ran to assess the whole system on both in-domain topics and various out-of-domain topics. The input data utilized for the evaluation consisted of 232 sentences that were fed into the system. The analysis was done in two steps. First, the evaluators judged whether and to what extent the NLU and conversational mover were successful in categorizing the input. Second, they looked at the system capability’s to produce a sensible and adequate response for those input sentences that were correctly classified by both the NLU and conversational mover in the previous processing step.

The analysis of the first step revealed that out of the 232 input sentences, the system was able to assign a correct label to 186 of them which amounts to 80.17% of the entire input data set. The system didn’t produce any label for some 21 sentences, i.e. 9.05% of the input data. The remaining 25 sentences, amounting to 10.78% of the whole data set, were instead given wrong labels and thus ultimately counts as incorrectly classified. When the system is not able to assign any label, HCA expresses its inability to address the input by giving an answer similar to answer 4 in Table 5. Table 4) shows the result of the label assignment task by the system as produced by the cascaded processing of NLU and conversational mover. There were full agreement between the evaluators on the categorization outcome.

Further, the evaluators focused on the 186 sentences that were correctly classified by the conversational mover. These sentences were then passed on to the system’s modules that resort to web resources and QA systems calls for the generation of a response to the user. A fine-grained classification scheme for the analysis of this task was agreed upon by the evaluators. Table 5 outlines such a scheme by providing a few explanatory inputs, one for each defined class. Interrater reliability for the response classification task was 98%. A third human judge opinion was used to resolve the problem cases. These are the four categories employed for response classification:

- **Right Answer**: This classification is used when the answer completely addresses the question, is correctly formulated and contains no grammatical errors.
- **Average Answer**: This class includes answers that still correctly address the user input but are either very short (typically this is the case of a semantically correct answer which is however pragmatically not correct) or grammatically incorrect. This class is also a container for responses that address the topic at hand but do not properly address the question. For instance, in Table 5, the question posed aims to know about the time when the radio was invented. The answer proposed does address the topic "radio" but refers to a specific kind of

<table>
<thead>
<tr>
<th>User</th>
<th>What do you think about Agatha Christie</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLU</td>
<td><a href="">dialogact:question</a><a href="">dialogactype:general</a><a href="">concept:unknown</a><a href="">subconcept:agatha_christie</a> <a href="">property:think</a></td>
</tr>
<tr>
<td>Google</td>
<td>Classification <a href="">dialogact:question</a><a href="">dialogactype:general</a><a href="">concept:unknown</a><a href="">subconcept:agatha_christie</a> <a href="">property:think</a></td>
</tr>
<tr>
<td>Conversational Mover</td>
<td>famous_personality_opinion</td>
</tr>
<tr>
<td>Web Agent</td>
<td>I think Agatha Christie is a fantastic author, and I am looking forward to reading another one of her books</td>
</tr>
</tbody>
</table>

Table 3: Output produced by a few system components for three examples taken from the log files recorded during evaluation.
The system was unable to handle properly. In order to understand them better, we describe them in the following with explanatory use cases.

**Use case 1:** In this example, the classification approach faces problems when the group of words overlaps with the words in the lexicon.

**Question:** Do you like the movie The Lord of the Rings?

**Answer:** I am sorry I don’t have an answer

In this case, the words "of" and "the" have a lexical entry, their category is retrieved from the lexicon and the only unknown words remaining are "Lord" and "Rings" and the web agent is not able to find the correct category for these individual words.

**Use case 2:** In this situation the system has difficulty in finding the right position to cut a long answer into smaller piece that could be sent to the TTS. Currently the system uses the strategy of picking the first two sentences, however, in some cases, a more appropriate answer exists in the middle two sentences.

**Question:** Did you know Oliver Twist?

**Answer:** Oliver Twist has been the subject of numerous film and television adaptations. It has been the basis for a highly successful musical Oliver.

A statistics for the four different output categories according to the response classification is shown in Table 6. The statistic judges the system on its ability to produce a response using web resources and QA systems and is based on the 186 sentences correctly classified by the conversational mover. The figures in that table indicate that a "positive answer" was provided in 84.40% of the cases (157 input sentences).

The evaluation also highlighted other problematic cases that the system was unable to handle properly. In order to understand them better, we describe them in the following with explanatory use cases.

**Use case 3:** In this last example, the system only finds a web site that could provide the right answer to the user question. The right answer has to be retrieved by traversing multiple links.

**Question:** What is the UEFA Champions League?

**Answer:** uefa.com - UEFA Champions League

In this case, a better answer would have been if the system would have picked the answer Oliver Twist is Charles Dickens’ second novel. The book was originally published in Bentley’s Miscellany as a serial. (approx the middle two sentences from the retrieved answer)

**Use case 4:** In this case the system is completely off-topic or non-sense. It has been the basis for a highly successful musical Oliver. The book was originally published in Bentley’s Miscellany as a serial. (approx the middle two sentences from the retrieved answer)

**Question:** Did you know Oliver Twist?

**Answer:** Oliver Twist has been the subject of numerous film and television adaptations. It has been the basis for a highly successful musical Oliver.

### Table 4: Input sentences labeling outcome.

<table>
<thead>
<tr>
<th></th>
<th>Right Label</th>
<th>No Label</th>
<th>Wrong Label</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># of sentences</td>
<td>186</td>
<td>21</td>
<td>25</td>
<td>232</td>
</tr>
<tr>
<td>Percentage</td>
<td>80.17%</td>
<td>9.05%</td>
<td>10.78%</td>
<td>100%</td>
</tr>
</tbody>
</table>

### Table 5: Classification of sentences according to the output they produced; the sentences taken into account are those correctly categorized by the conversational mover.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Question/Answer Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right Answer</td>
<td>who was Mozart</td>
</tr>
<tr>
<td></td>
<td>Mozart is one of the heavyweights of classical music, generally placed in the top rank of composers along with Beethoven and Bach</td>
</tr>
<tr>
<td>Average Answer</td>
<td>when was invented the radio</td>
</tr>
<tr>
<td></td>
<td>Car Radio In 1929 American Paul Galvin the head of Galvin Manufacturing Corporation invented the first car radio</td>
</tr>
<tr>
<td>Wrong Answer</td>
<td>when was the second world war</td>
</tr>
<tr>
<td></td>
<td>She was launched in March 1938 and served throughout the Second World War playing a leading part in the destruction of the German battle cruiser</td>
</tr>
<tr>
<td>No Answer</td>
<td>Olympic games are coming soon</td>
</tr>
<tr>
<td></td>
<td>I am sorry I don’t have an answer</td>
</tr>
</tbody>
</table>

8. DISCUSSION

Existing work on ECAs development has mainly centered on task oriented systems where the goal of the system is to cooperate with the human participant to solve a task within a given domain effectively and efficiently. Their main limitation, its fixed context, is simultaneously its greatest strength since it allows building very robust and feasible systems. However, they are a simplification of real human conversational behavior, for they control and restrict the interaction rather than enrich it. As we start developing systems that provide a rich social interaction, we need strategies for expanding on the range of conversation topics. These discussion topics could range from multiple general-purpose subjects and cover the most disparate topics. The simplest way
to address an unknown topic would be either a standard reply "I don’t know", ignoring it altogether or through shallow pattern matching techniques as used by chatterbots.

Our approach is based on a clever idea which, like most good ideas, is a pretty simple one. For language understanding purposes related to these topics, we have employed an approach that exploits Google's directory structure along with existing domain independent properties and dialog acts to build a consistent representation with other domain input. There have been related approaches that uses Yahoo's categories [12] for document classification of documents using an N-gram classifier. We are not aware of any approaches utilizing directory categorization for language understanding. Moreover we handle out-of-domain topics by resorting to freely available QA systems and data posted on web sites. There have been attempts to use dialog strategies for a better performance of QA systems by using clarification dialogs to refine user questions [5] but we do not know of any attempts to use the ability to address various topics provided by QA system along with the normal domain oriented conversation. Knowledge available on web sites has traditionally been utilized to develop three types of dialog systems: Information Retrieval aid dialog, form-filling dialog and a table-based dialog [9]. However, the purpose of these dialog systems is to help the user use the web site efficiently. Thus essentially, it acts as a front end to the web site. Our approach utilizes the data posted on the web to converse on general purpose topics apart from a discussion on domain oriented topics.

The evaluation indicates that our system can generate a correct response to a user turn in 67.67% of the cases. The evaluation also highlights certain drawbacks of the system that need to be addressed in the future. One drawback is that the classification approach faces problems when the group of words overlaps with the words in the lexicon. One solution would be to automatically detect the entries, which overlap with the words in the lexicon by parsing the Google's directory structure off-line and having these entries made in the key phrase spotter. This issue however arises only when important keywords are removed from the names of movie, game and/or famous personalities due to lexicon overlap. For instance, in example 3 of Table 3, despite the word "my" was removed, the Google’s directory structure is still able to provide an appropriate categorization as sufficient keywords are present.

The evaluation also brought to surface the problem of finding the positions within a generated long sentence where to split it into a sequence of shorter ones. We are also still trying to find a solution to the problem of traversing hyper-links to find the right answer from a web page. An additional limitation of our current implementation is that once a general unknown topic is identified and a QA system is called to generate a response, the dialog is suspended. This is a critical feature, because in some sense this means that at times the system performs a QA cycle rather than a dialog. We had to follow this strategy to prevent the user starting a dialog about the unknown topic which would result in a series of QA cycles only. In fact, the conversational character is not able to continue a contextual conversational exchange beyond the sentence he selects as reply on unknown topics. Having the knowledge of topics addressable through QA systems and web sites, we aim to conduct a mixed initiative dialog on these topics along with the normal domain-oriented conversation. At this stage, it has not been possible to us to give the character a personalized attitude and beliefs towards these general-purpose topics. We aim to provide more life-like behaviors by resolving these issues. The system parsing the data available on QA system and web sites is susceptible to changes in the format of web sites. We have considered this formatting issues, and in fact we have designed our parsing algorithm in a way that the virtual agent does not utter any non-sense such as html instructions, xml texts, control characters etc.

9. CONCLUSION AND FUTURE WORK

The objective of developing conversational agents capable of addressing general purpose everyday topics is a very difficult task. Such an ambitious goal is paramount to the success and acceptance of any computing systems populated by interactive characters. In principle, conversational characters should not be expected to conduct a simulated conversation that exclusively revolves around their domains of expertise. They should be endowed with the capability of reaching out into topics that could not be covered by the developers during the creation of the system. Hence, one of the future research efforts in developing a compelling experiential interaction with embodied agents requires not only the expansion of the repertoire of discussion topics but also effective strategies for handling out-of-domain input. Unfortunately, creating resources for agents with ability to address these topics from scratch is a major endeavor.

In this paper, we have presented an approach to handle the general purpose topics by using data available from web sites and freely accessible QA systems. It reduces content creation authorial burden for these topics and provides conversational agents with the ability to talk about general purpose topics. On the understanding side, we have used Google’s directory classification mechanism along with existing domain independent dialog acts and properties to understand these general purpose topics. Automatic categorization provided by Google’s directory allows for easy and fast addition of these general purpose topics.

At this point, our research prototype is sufficiently stable and the evaluation shows promising results. Our results indicate a proper retrieval of a reply in presence of out of domain topics in 68% of the cases analyzed. This first evaluation also gave us helpful hints about improvements that need to be carried out in newer versions of the system. System acceptance may be dependent on or impeded by several factors, such a task takes a long time. Hence, we have been looking at a variety of quantitative measures (e.g. accuracy, time to perform, success rate, percent agreement of assessments) as well as qualitative ones (e.g. user perceptions of utility, ease of use, mental and temporal demands, user frustration and naturalness) that should we plan to measure in our upcoming usability testing sessions. We are particularly interested in pursuing a strategy adopted in [22] to score conversational systems with a combination of dialog success measure and various utterance-related costs as well as approaches like [17] to measure improvements in user experience through our techniques.
<table>
<thead>
<tr>
<th>Sentence Type</th>
<th>Total Answer</th>
<th>Right Answer</th>
<th>Average Answer</th>
<th>Wrong Answer</th>
<th>No Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>When question</td>
<td>81</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Who question</td>
<td>36</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>What question</td>
<td>65</td>
<td>49</td>
<td>8</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Statements</td>
<td>91</td>
<td>55</td>
<td>10</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Miscellaneous/Other</td>
<td>26</td>
<td>9</td>
<td>12</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>186</td>
<td>101</td>
<td>56</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>Percentage</td>
<td>100%</td>
<td>54.30%</td>
<td>30.1%</td>
<td>6.45%</td>
<td>9.15%</td>
</tr>
</tbody>
</table>

Table 6: Statistics of the responses based on the classification scheme chosen by the evaluators; the sentences used for the analysis are the ones that are correctly classified by both NLU and Conversational Mover.

10. REFERENCES