Demonstration of a prototype for a Conversational Companion for reminiscing about images

Yorick Wilks
IHMC, Florida
ywilks@ihmc.us
Alexiei Dingli
University of Malta, Malta
alexiei.dingli@um.edu.mt
Roberta Catizone
University of Sheffield, UK
r.catizone@dcs.shef.ac.uk
Weiwei Cheng
University of Sheffield, UK
w.cheng@dcs.shef.ac.uk

Abstract

This paper describes an initial prototype demonstrator of a Companion, designed as a platform for novel approaches to the following: 1) The use of Information Extraction (IE) techniques to extract the content of incoming dialogue utterances after an Automatic Speech Recognition (ASR) phase, 2) The conversion of the input to Resource Descriptor Format (RDF) to allow the generation of new facts from existing ones, under the control of a Dialogue Manger (DM), that also has access to stored knowledge and to open knowledge accessed in real time from the web, all in RDF form, 3) A DM implemented as a stack and network virtual machine that models mixed initiative in dialogue control, and 4) A tuned dialogue act detector based on corpus evidence. The prototype platform was evaluated, and we describe this briefly; it is also designed to support more extensive forms of emotion detection carried by both speech and lexical content, as well as extended forms of machine learning.

1. Introduction

This demonstrator Senior Companion (SC) was built during the initial phase of the Companions project and aims to change the way we think about the relationships of people to computers and the internet by developing a virtual conversational 'Companion that will be an agent or 'presence' that stays with the user for long periods of time, developing a relationship and 'knowing its owners' preferences and wishes. The Companion communicates with the user primarily through speech, but also using other technologies such as touch screens and sensors.

This paper describes the functionality and system modules of the Senior Companion, one of two initial prototypes built in the first two years of the project. The SC provides a multimodal interface for eliciting, retrieving and inferring personal information from elderly users by means of conversation about their photographs. The Companion, through conversation, elicits life memories and reminiscences, often prompted by discussion of their photographs; the aim is that the Companion should come to know a great deal about its user, their tastes, likes, dislikes, emotional reactions etc, through long periods of conversation. It is assumed that most life information will soon be stored on the internet (as in the Memories for Life project: http://www.memoriesforlife.org/) and we have linked the SC directly to photo inventories in Facebook (see below). The overall aim of the SC project (not yet achieved) is to produce a coherent life narrative for its user from conversations about personal photos, although its short-term goals, reported here, are to assist, amuse and entertain the user.

The technical content of the project is to use a number of types of machine learning (ML) to achieve these ends in original ways, initially using a methodology developed in earlier research: first, by means of an Information Extraction (IE) approach to deriving content from user input utterances; secondly, using a training method for attaching Dialogue Acts to these utterance and, lastly, using a specific type of dialogue manager (DM) that uses Dialogue Action Forms (DAF) to determine the context of any utterance. A stack of these DAFs is the virtual machine that models the ongoing dialogue by means of shared user and Companion initiative and generates appropriate responses. In this description of the demo, we shall:

- describe the current SC prototype's functionality;
- set out its architecture and modules, focusing on the Natural Language Understanding module and the Dialogue Manager.

A mini-version of the demo running in real time can be seen at: URL

Proceedings of the ACL 2010 System Demonstrations, pages 72–77,
Uppsala, Sweden, 13 July 2010.

http://www.youtube.com/watch?v=-Xx5hgjD-Mw

2. The Senior Companion System

The Senior Companion prototype (Wilks, 2010) was designed to make a rapid advance in the first two years of the project so as to be basis for a second round of prototypes embodying more advanced ML. This strategy was deliberately chosen to avoid a well-known problem with experimental AI systems: that a whole project is spent in design so that a prototype never emerges until the very end, but which is then never fully evaluated and, most importantly, nothing is ever built on the experience obtained in its construction. The central function of the SC is engaging the user in discussion about their photographs: where and when they were taken, details about the people in them and their relationship to the user and each other.

Once a photo is loaded, it is processed with face recognition software to identify any faces in it. The recognition software, OpenCV, provides positional information by identifying the face coordinates and this information is exploited in the Dialogue Manager by making explicit reference to the position of people in the photograph (the person on the left, right, center, etc.) as well as recognizing when there are groups of people. The system discusses properties of the photo as well as properties and relationships of the people in the photos.

The news reading feature adds an interesting accompaniment to the photo domain and demonstrates the ability of the system to handle more than one kind of application at a time, and news has, of course, an unconstrained vocabulary. The following is a fairly typical example of its current capacity, depending of course on the images loaded, and comes from the middle part of a sample dialogue generated by the system when discussing a group photo from a vacation in Zanzibar, and after the system had already stored, from an earlier conversation, that the user has a daughter Zoe (using researchers' children to avoid any problems):

SC: What is your relationship to the doctor?
User: There is no relationship.
SC: Please tell me how you know the doctor
U: Uh, he's a friend.
SC: How old was the doctor when this photo was taken?
U: I'm not sure but I think about 30.
SC: How do you know Octavia?
U: She's my daughter.
SC: Aha! I see. Octavia's sister is Zoe.
SC: How old was your daughter Octavia when this photo was taken?
U: Octavia was 19 years old.
SC: What was the occasion of this photo?
U: This was another family holiday.
SC: Where was this photo taken?
U: This photo was taken in Tanzania.
SC: have you ever been to Zanzibar?
U: Yes, I have.
SC: Great place, I hope you saw the beaches. When was this photo taken?
U: last year.

It is a key feature of the SC that it can make simple inferences from family relationships it already knows (e.g. that daughters of the same parent are siblings) and that it can access real-time information about places to show that it has some knowledge of what is being talked about, in this case the beaches on Zanzibar, and how this is done is discussed below. This real time access to unconstrained place information on the internet is an attempt to break out of classic AI systems that only know the budget of facts they have been primed with.

This basic system provides the components for future development of the SC, as well as its main use as a device to generate more conversation data for machine learning research in the future. Key features of the SC are listed below followed by a description of the system architecture and modules. The SC:

- Contains a visually appealing multimodal interface with a character avatar to mediate the system's functionality to the user.
- Interacts with the user using multiple modalities – speech and touch.
- Includes face detection software for identifying the position of faces in the photos.
- Accepts pre-annotated (XML) photo inventories as a means for creating richer dialogues more quickly.
- Engages in conversation with the user about topics within the photo domain: when and where the photo was taken, discussion of the people in the photo including their relationships to the user.
- Reads news from three categories: politics, business and sports.

- Tells jokes taken from an internet-based joke website.
- Retains all user input for reference in repeat user sessions, in addition to the knowledge base that has been updated by the Dialogue Manager on the basis of what was said.
- Contains a fully integrated Knowledge Base for maintaining user information including:
    - Ontological information which is exploited by the Dialogue Manager and provides domain-specific relations between fundamental concepts.
    - A mechanism for storing information in a triple store (Subject-Predicate-Object) - the RDF Semantic Web format - for handling unexpected user input that falls outside of the photo domain, e.g. arbitrary locations in which photos might have been taken.
    - A reasoning module for reasoning over the Knowledge Base and world knowledge obtained in RDF format from the internet; the SC is thus a primitive Semantic Web device (see refernce8, 2008)
- Contains basic photo management capability allowing the user, in conversation, to select photos as well as display a set of photos with a particular feature.

Figure 1: The Senior Companion Interface

3. System Architecture

In this section we will review the components of the SC architecture. As can be seen from Figure 2, the architecture contains three abstract level components – Connectors, Input Handlers and Application Services –together with the Dialogue Manager and the Natural Language Understander (NLU).

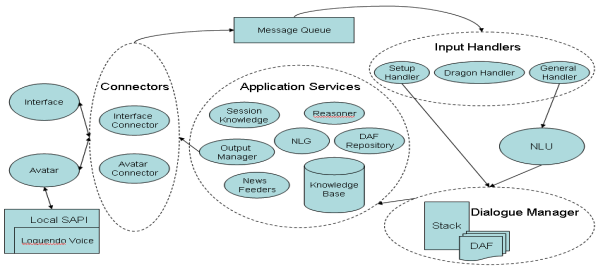

Figure 2: Senior Companion system architecture

Connectors form a communication bridge between the core system and external applications. The external application refers to any modules or systems which provide a specific set of functionalities that might be changed in the future. There is one connector for each external application. It hides the underlying complex communication protocol details and provides a general interface for the main system to use. This abstraction decouples the connection of external and internal modules and makes changing and adding new external modules easier. At this moment, there are two connectors in the system – Napier Interface Connector and CrazyTalk Avatar Connector. Both of them are using network sockets to send/receive messages.

Input Handlers are a set of modules for processing messages according to message types. Each handler deals with a category of messages where categories are coarse-grained and could include one or more message types. The handlers separate the code handling inputs into different places and make the code easier to locate and change. Three handlers have been implemented in the Senior Companion system – Setup Handler, Dragon Events Handler and General Handler. The Setup Handler is responsible for loading the photo annotations if any, performing face detection if no annotation file is associated with the photo and checking the Knowledge Base in case

the photo being processed has been discussed in earlier sessions. Dragon Event Handler deals with dragon speech recognition commands sent from the interface while the General Handler processes user utterances and photo change events of the interface.

Application Services are a group of internal modules which provide interfaces for the Dialogue Action Forms (DAF) to use. It has an easy-to-use high-level interface for general DAF designers to code associated tests and actions as well as a low level interface for advanced DAFs. It also provides the communication link between DAFs and the internal system and enables DAFs to access system functionalities. Following is a brief summary of modules grouped into Application Services.

News Feeders are a set of RSS Feeders for fetching news from the internet. Three different news feeders have been implemented for fetching news from BBC website Sports, Politics and Business channels. There is also a Jokes Feeder to fetch Jokes from internet in a similar way. During the conversation, the user can request news about particular topics and the SC simply reads the news downloaded through the feeds.

The DAF Repository is a list of DAFs loaded from files generated by the DAF Editor.

The Natural Language Generation (NLG) module is responsible for randomly selecting a system utterance from a template. An optional variable can be passed when calling methods on this module. The variable will be used to replace special symbols in the text template if applicable.

Session Knowledge is the place where global information for a particular running session is stored. For example, the name of the user who is running the session, the list of photos being discussed in this session and the list of user utterances etc.

The Knowledge Base is the data store of persistent knowledge. It is implemented as an RDF triplestore using a Jena implementation. The triplestore API is a layer built upon a traditional relational database. The application can save/retrieve information as RDF triples rather than table records. The structure of knowledge represented in RDF triples is discussed later.

The Reasoner is used to perform inference on existing knowledge in the Knowledge Base (see example in next section).

The Output Manager deals with sending messages to external applications. It has been implemented in a publisher/subscriber fashion. There are three different channels in the system: the text channel, the interface command channel and the avatar command channel. Those channels could be subscribed to by any connectors and handled respectively.

4. Dialogue understanding and inference

Every utterance is passed through the Natural Language Understanding (NLU) module for processing. This module uses a set of well-established natural language processing tools such as those found in the GATE (Cunningham, et al., 1997) system. The basic processes carried out by GATE are: tokenizing, sentence splitting, POS tagging, parsing and Named Entity Recognition. These components have been further enhanced for the SC system by adding 1) new and improved gazetteers including family relations and 2) accompanying extraction rules .The Named Entity (NE) recognizer is a key part of the NLU module and recognizes the significant entities required to process dialogue in the photo domain: PERSON NAMES, LOCATION NAMES, FAMILY RELATIONS and DATES. Although GATE recognizes basic entities, more complex entities are not handled. Apart from the gazetteers mentioned earlier and the hundreds of extraction rules already present in GATE, about 20 new extraction rules using the JAPE rule language were also developed for the SC module. These included rules which identify complex dates, family relationships, negations and other information related to the SC domain. The following is an example of a simple rule used to identify relationship in utterances such as “Mary is my sister”:

```
Macro: RELATIONSHIP_IDENTIFIER
(
({To-
ken.category=="PRP$"}|{Token.category=="PR
P"}|{Lookup.majorType=="person_first"}):pers
on2
({Token.string=="is"})
({Token.string=="my"}):person1
({Lookup.minorType=="Relationship"}):relation
ship)
```

Using this rule with the example mentioned earlier, the rule interprets person1 as referring to the speaker so, if the name of the user speaking is John (which was known from previous conversations), it is utilized. Person 2 is then the name of the person mentioned, i.e. Mary. This name is recognised by using the gazetteers we have in the system (which contain about 40,000 first names). The relationship is once again identified using the almost 800 unique relationships added to the gazetteer. With this information, the NLU module identifies Information Extraction patterns in the dialogue that represent significant content with respect to a user's life and photos.

The information obtained (such as Mary=sister-of John) is passed to the Dialogue Manager (DM) and then stored in the knowledge base (KB). The DM filters what to include and exclude from the KB. Given, in the example above, that Mary is the sister of John, the NLU knows that sister is a relationship between two people and is a key relationship. However, the NLU also discovers syntactical information such as the fact the both Mary and John are nouns. Even though this information is important, it is too low level to be of any use by the SC with respect to the user, i.e. the user is not interested in the parts-of-speech of a word. Thus, this information is discarded by the DM and not stored in the KB. The NLU module also identifies a Dialogue Act Tag for each user utterance based on the DAMSL set of DA tags and prior work done jointly with the University of Albany (Webb et al., 2008).

The KB is a long-term store of information which makes it possible for the SC to retrieve information stored between different sessions. The information can be accessed anytime it is needed by simply invoking the relevant calls. The structure of the data in the database is an RDF triple, and the KB is more commonly referred to as a triple store. In mathematical terms, a triple store is nothing more than a large database of interconnected graphs. Each triple is made up of a subject, a predicate and an object. So, if we took the previous example, Mary sister-of John; Mary would be the subject, sister-of would be the predicate and John would be the object. The inference engine is an important part of the system because it allows us to discover new facts beyond what is elicited from the conversation with the user.

Uncle Inference Rule:
(?a sisterOf ?b),
(?x sonOf ?a),
(?b gender male) -> (?b uncleOf ?x)

Triples:
(Mary sisterOf John)
(Tom sonOf Mary)

Triples produced automatically by ANNIE (the semantic tagger):
(John gender male)

Inference:
(Mary sisterOf John)
(Tom sonOf Mary)
(John gender male)
->
(John uncleOf Tom)

This kind of inference is already used by the SC and we have about 50 inference rules aimed at producing new data on the relationships domain. This combination of triple store, inference engine and inference rules makes a system which is weak but powerful enough to mimic human reasoning in this domain and thus simulate basic intelligence in the SC. For our prototype, we are using the JENA Semantic Web Framework for the inference engine together with a MySQL database as the knowledge base. However, this system of family relationships is not enough to cover all the possible topics which can crop up during a conversation and, in such circumstances, the DM switches to an open-world model and instructs the NLU to seek further information online.

5. The Hybrid-world approach

When the DM requests further information on a particular topic, the NLU first checks with the KB whether the topic is about something known. At this stage, we have to keep in mind that any topic requested by the DM should be already in the KB since it was preprocessed by the NLU when it was mentioned in the utterance. So, if the user informs the system that the photograph was taken in Paris, (in response to a system question asking where the photo was taken), the utterance is first processed by the NLU which discovers that “Paris” is a location using its semantic tagger ANNIE (A Nearly New Information Extraction engine). The semantic tagger makes use of gazetteers and IE rules in order to accomplish

this task. It also goes through the KB and retrieves any triples related to "Paris". Inference is then performed on this data and the new information generated by this process is stored back in the KB.

Once the type of information is identified, the NLU can use various predefined strategies: In the case of LOCATIONS, one of the strategies used is to seek for information in Wiki-Travel or Virtual Tourists. The system already knows how to query these sites and interpret their output by using predefined wrappers. This is then used to extract relevant information from the mentioned sites webpages by sending an online query to these sites and storing the information retrieved in the triple-store. This information is then used by the DM to generate a reply. In the previous example, the system manages to extract the best sightseeing spots in Paris. The NLU would then store in the KB triples such as [Paris, sightseeing, Eiffel Tower] and the DM with the help of the NLG would ask the user "I've heard that the X is a very famous spot. Have you seen it while you were there?" Obviously in this case, X would be replaced by the "Eiffel Tower".

On the other hand, if the topic requested by the DM is unknown, or the semantic tagger is not capable of understanding the semantic category, the system uses a normal search engine (and this is what we call "hybrid-world": the move outside the world the system already knows). A query containing the unknown term in context is sent to standard engines and the top pages are retrieved. These pages are then processed using ANNIE and their tagged attributes are analyzed. The standard attributes returned by ANNIE include information about Dialogue Acts, Polarity (i.e. whether a sentence has positive, negative or neutral connotations), Named Entities, Semantic Categories (such as dates and currency), etc. The system then filters the information collected by using more generic patterns and generates a reply from the resultant information. ANNIE's polarity methods have been shown to be an adequate implementation of the general word-based polarity methods pioneered by Wiebe and her colleagues (see e.g. Akkaya et al., 2009).

6. Evaluation

The notion of companionship is not yet one with any agreed evaluation strategy or metric, though developing one is part of the main project itself. Again, there are established measures for the assessment of dialogue programs but they have all been developed for standard task-based dialogues and the SC is not of that type: there is no specific task either in reminiscing conversations, nor in the elicitation of the content of photos, that can be assessed in standard ways, since there is no clear point at which an informal dialogue need stop, having been completed. Conventional dialogue evaluations often use measures like "stickiness" to determine how much a user will stay with or stick with a dialogue system and not leave it, presumably because they are disappointed or find it lacking in some feature. But it is hard to separate that feature out from a task rapidly and effectively completed, where stickiness would be low not high. Traum (Traum et al., 2004) has developed a methodology for dialogue evaluation based on "appropriateness" of responses and the Companions project has developed a model of evaluation for the SC based on that (Benyon et al., 2008).

Acknowledgement

This work was funded by the Companions project (2006-2009) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434.

References

David Benyon, Prem Hansen and Nick Webb, 2008. Evaluating Human-Computer Conversation in Companions. In: *Proc.4th International Workshop on Human-Computer Conversation*, Bellagio, Italy.

Cem Akkaya, Jan Wiebe, and Rada Mihalcea,. 2009. Subjectivity Word Sense Disambiguation, In: *EMNLP 2009.*

Hamish Cunningham, Kevin Humphreys, Robert Gaizauskas, and Yorick Wilks, 1997. GATE -- a TIPSTER based General Architecture for Text Engineering. In: *Proceedings of the TIPSTER Text Program (Phase III) 6 Month Workshop*. Morgan Kaufmann, CA.

David Traum, Susan Robinson, and Jens Stephan. 2004. Evaluation of multi-party virtual reality dialogue interaction, In: *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pp.1699-1702

Yorick Wilks (ed.) 2010. *Artificial Companions in Society: scientific, economic, psychological and philosophical perspectives.* John Benjamins: Amsterdam.