Meet EDGAR, a tutoring agent at MONSERRATE

Pedro Fialho, Luísa Coheur, Sérgio Curto, Pedro Cláudio
Ângela Costa, Alberto Abad, Hugo Meinedo and Isabel Trancoso
Spoken Language Systems Lab (L2F), INESC-ID
Rua Alves Redol 9
1000-029 Lisbon, Portugal
name.surname@l2f.inesc-id.pt

Abstract

In this paper we describe a platform for embodied conversational agents with tutoring goals, which takes as input written and spoken questions and outputs answers in both forms. The platform is developed within a game environment, and currently allows speech recognition and synthesis in Portuguese, English and Spanish. In this paper we focus on its understanding component that supports in-domain interactions, and also small talk. Most in-domain interactions are answered using different similarity metrics, which compare the perceived utterances with questions/sentences in the agent's knowledge base; small-talk capabilities are mainly due to AIML, a language largely used by the chatbots' community. In this paper we also introduce EDGAR, the butler of MONSERRATE, which was developed in the aforementioned platform, and that answers tourists' questions about MONSERRATE.

1 Introduction

Several initiatives have been taking place in the last years, targeting the concept of Edutainment, that is, education through entertainment. Following this strategy, virtual characters have animated several museums all over the world: the 3D animated Hans Christian Andersen is capable of establishing multimodal conversations about the writer's life and tales (Bernsen and Dybkjr, 2005), Max is a virtual character employed as guide in the Heinz Nixdorf Museums Forum (Pfeiffer et al., 2011), and Sergeant Blackwell, installed in the Cooper-Hewitt National Design Museum in New York, is used by the U.S. Army Recruiting Command as a hi-tech attraction and information source (Robinson et al., 2008). DuARTE Digital (Mendes et al., 2009) and EDGAR are also examples of virtual characters for the Portuguese language with the same edutainment goal: DuARTE Digital answers questions about Custódia de Belém, a famous work of the Portuguese jewelry; EDGAR is a virtual butler that answers questions about MONSERRATE (Figure 1).

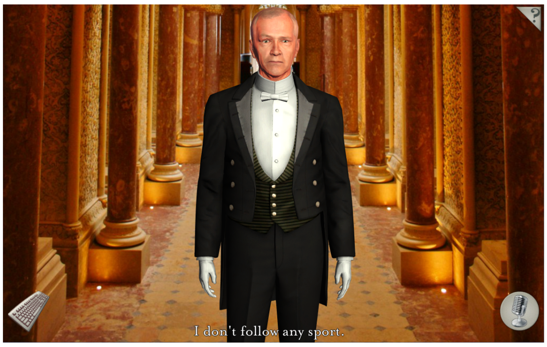

I don't follow any sport.

Figure 1: EDGAR at MONSERRATE.

Considering the previous mentioned agents, they all cover a specific domain of knowledge (although a general Question/Answering system was integrated in Max (Waltinger et al., 2011)). However, as expected, people tend also to make small talk when interacting with these agents. Therefore, it is important that these systems properly deal with it. Several strategies are envisaged to this end and EDGAR is of no exception. In this paper, we describe the platform behind EDGAR, which we developed aiming at the fast insertion of in-domain knowledge, and to deal with small talk. This platform is currently in the process of being industrially applied by a company known for its expertise in building and deploying kiosks. We will provide the hardware and software required to demonstrate EDGAR, both on a computer and on a tablet.

This paper is organized as follows: in Section 2 we present EDGAR's development platform

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 61–66,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

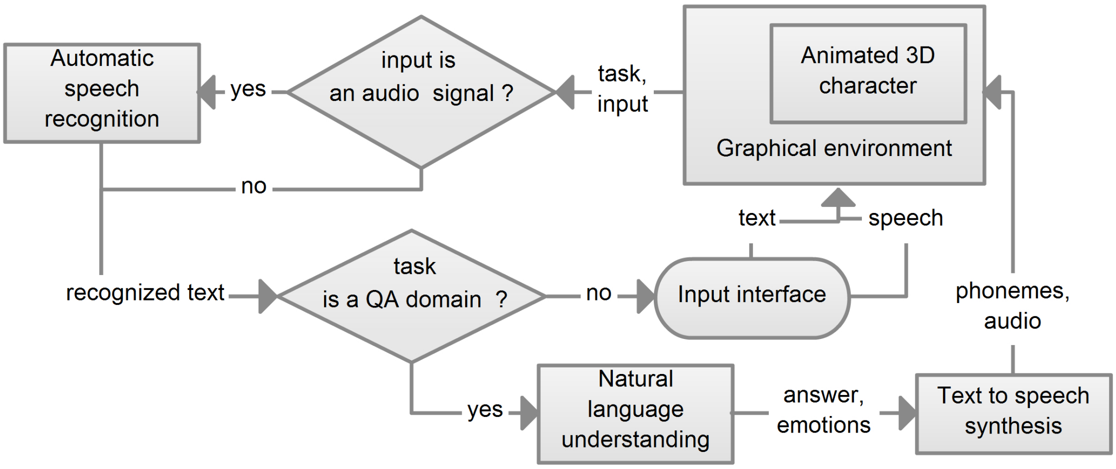


Figure 2: EDGAR architecture

and describe typical interactions, in Section 3 we show how we move from in-domain interactions to small talk, and in Section 4 we present an analysis on collected logs and their initial evaluation results. Finally, in Section 5 we present some conclusions and point to future work.

2 The Embodied Conversational Agent platform

2.1 Architecture overview

The architecture of the platform, generally designed for the development of Embodied Conversational Agents (ECAs) (such as EDGAR), is shown in Figure 2. In this platform, several modules intercommunicate by means of well defined protocols, thus leveraging the capabilities of independent modules focused on specific tasks, such as speech recognition or 3D rendering/animation. This independence allows us to use subsets of this platform modules in scenarios with different requirements (for instance, we can record characters uttering a text).

Design and deployment of the front end of EDGAR is performed in a game engine, which has enabled the use of computer graphics technologies and high quality assets, as seen in the video game industry.

2.2 Multimodal components

The game environment, where all the interaction with EDGAR takes place, is developed in the Unity[1] platform, being composed of one highly detailed character, made and animated by Rocketbox studios[2], a virtual keyboard and a push-while-talking button.

In this platform, Automatic Speech Recognition (ASR) is performed by AUDIMUS (Meinedo et al., 2003) for all languages, using generic acoustic and language models, recently compiled from broadcast news data (Meinedo et al., 2010). Language models were interpolated with all the domain questions defined in the Natural Language Understanding (NLU) framework (see below), while ASR includes features such as speech/non-speech (SNS) detection and automatic gain control (AGC). Speech captured in a public space raises several ASR robustness issues, such as loudness variability of spoken utterances, which is particularly bound to happen in a museological environment (such as MONSERRATE) where silence is usually incited. Thus, we have added a bounded amplication to the captured signal, despite the AGC mechanism, ensuring that too silent sounds are not discarded by the SNS mechanism.

Upon a spoken input, AUDIMUS translates it into a sentence, with a confidence value. An empty recognition result, or one with low confidence, triggers a control tag ("_REPEAT_") to the NLU module, which results in a request for the user to repeat what was said. The answer returned by the NLU module is synthesized in a language dependent Text To Speech (TTS) system, with DIXI (Paulo et al., 2008) being used for Portuguese, while a recent version of FESTIVAL (Zen et al., 2009) covers both English and Spanish. The

[1] http://unity3d.com/

[2] http://www.rocketbox-libraries.com/

synthesized audio is played while the corresponding phonemes are mapped into visemes, represented as skeletal animations, being synchronized according to phoneme durations, available in all the employed TTS engines.

Emotions are declared in the knowledge sources of the agent. As shown in Figure 3, they are coordinated with viseme animations.

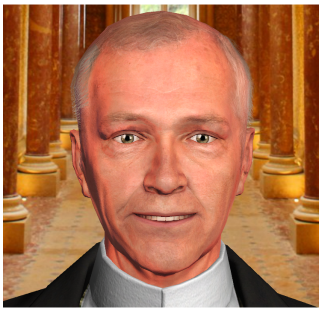
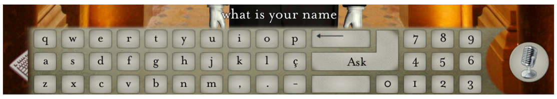


Figure 3: The EDGAR character in a joyful state.

2.3 Interacting with EDGAR

In a typical interaction, the user enters a question with a virtual keyboard or says it to the microphone while pressing a button (Figure 4), in the language chosen in the interface (as previously said, Portuguese, English or Spanish).



what is your name

Figure 4: A question written in the EDGAR interface.

Then, the ASR will transcribe it and the NLU module will process it. Afterwards, the answer, chosen by the NLU module, is heard through the speakers, due to the TTS, and sequentially written in a talk bubble, according to the produced speech. The answer is accompanied with visemes, represented by movements of the character's mouth/lips, and by facial emotions as marked in the answers of the NLU knowledge base. A demo of EDGAR, only for English interactions, can be tested in `https://edgar.l2f.inesc-id.pt/m3/edgar.php`.

3 The natural language understanding component

3.1 In-domain knowledge sources

The in-domain knowledge sources of the agent are XML files, hand-crafted by domain experts. This XML files have multilingual pairs constituted by different paraphrases of the same question and possible answers. The main reason to follow this approach (and contrary to other works where grammars are used), is to ease the process of creating/enriching the knowledge sources of the agent being developed, which is typically done by non experts in linguistics or computer science. Thus, we opted for following a similar approach of the work described, for instance, in (Leuski et al., 2006), where the agents knowledge sources are easy to create and maintain. An example of a questions/answers pair is:

```
<questions>
 <q en="How is everything?"
      es="Todo bien?">
       Tudo bem?</q>
</questions>
<answers>
  <a en="I am ok, thank you."
  es="Estoy bien, gracias."
  emotion="smile_02">
  Estou bem, obrigado.</a>
</answers>
```

As it can been see from this example, emotions are defined in these files, associated to each question/answer pair (emotion="smile" in the example, one of the possible smile emotions).

These knowledge sources can be (automatically) extended with "synonyms". We call them "synonyms", because they do not necessarily fit in the usual definition of synonyms. Here we follow a broader approach to this concept and if two words, within the context of a sentence from the knowledge source, will lead to the same answer, then we consider them to be "synonyms". For instance "palace" or "castle" are not synonyms. However, people tend to refer to MONSERRATE in both forms. Thus, we consider them to be "synonyms" and if one of these is used in the original knowledge sources, the other is used to expand them. It should be clear that we will generate many incorrect questions with this procedure, but empirical tests (out of the scope of this paper) show that these questions do not hurt the system performance. Moreover, they are useful for ASR language model interpolation, which is based on N-grams.

3.2 Out-of-domain knowledge sources

The same format of the previously described knowledge sources can be used to represent out-of-domain knowledge. Here, we extensively used the "synonyms" approach. For instance, words *wife* and *girlfriend* are considered to be "synonyms" as all the personal questions with these words should be answered with the same sentence: *I do not want to talk about my private life.*

Nevertheless, and taking into consideration the work around small talk developed by the chatbots community (Klwer, 2011), we decided to use the most popular language to build chatbots: the "Artificial Intelligence Markup Language", widely known as AIML, a derivative of XML. With AIML, knowledge is coded as a set of rules that will match the user input, associated with templates, the generators of the output. A detailed description of AIML syntax can be found in `http://www.alicebot.org/aiml.html`. In what respects AIML interpreters, we opted to use Program D (java), which we integrated in our platform. Currently, we use AIML to deal with slang and to answer questions that have to do with cinema and compliments.

As a curiosity, we should explain that we deal with slang when input came from the keyboard, and not when it is speech, as the language models are not trained with this specific lexicon. The reason we do that is because if the language models were trained with slang, it would be possible to erroneously detect it in utterances and then answer them accordingly, which could be extremely unpleasant. Therefore, EDGAR only deals with slang when the input is the keyboard.

The current knowledge sources have 152 question/answer pairs, corresponding to 763 questions and 206 answers. For Portuguese, English and Spanish the use of 226, 219 and 53 synonym relations, led to the generation of 22 194, 16 378 and 1 716 new questions, respectively.

3.3 Finding the appropriate answer

The NLU module is responsible for the answer selection process. It has three main components.

The first one, STRATEGIES, is responsible to choose an appropriate answer to the received interaction. Several strategies are implemented, including the ones based on string matching, string distances (as for instance, Levenshtein, Jaccard and Dice), N-gram Overlap and support vector machines (seeing the answer selection as a classification problem). Currently, best results are attained using a combination of Jaccard and bigram Overlap measures and word weight through the use of tf-idf statistic. In this case, Jaccard takes into account how many words are shared between the user's interaction and the knowledge source entry, bigram Overlap gives preference to the shared sequences of words and tf-idf contributes to the results attained by previous measures, by given weight to unfrequent words, which should have more weight on the decision process (for example, the word MONSERRATE occurs in the majority of the questions in the corpus, so it is not very informative and should not have the same weight as, for instance, the word *architect* or *owner*).

The second component, PLUGINS, deals with two different situations. First, it accesses Program D when interactions are not answered by the STRATEGIES component. That is, when the technique used by STRATEGIES returns a value that is lower than a threshold (dependent of the used technique), the PLUGIN component runs Program D in order to try to find an answer to the posed question. Secondly, when the ASR has no confidence of the attained transcription (and returns the "_REPEAT_" tag) or Program D is not able to find an answer, the PLUGINS component does the following (with the goal of taking the user again to the agent topic of expertise):

- In the first time that this occurs, a sentence such as *Sorry, I did not understand you.* is chosen as the answer to be returned.

- The second time this occurs, EDGAR asks the user *I did not understand you again. Why don't you ask me X?*, being X generated in run time and being a question from a subset of the questions from the knowledge sources. Obviously, only in-domain (not expanded) questions are considered for replacing X.

- The third time there is a misunderstanding, EDGAR says *We are not understanding each other, let me talk about* MONSERRATE. And it randomly choses some answer to present to the user.

The third component is the HISTORY-TRACKER, which handles the agent knowledge about previous interactions (kept until a default time without interactions is reached).

4 Preliminary evaluation

Edgar is more a domain-specific Question Answering (QA) than a task-oriented dialogue system. Therefore, we evaluated it with the metrics typically used in QA. The mapping of the different situations in true/false positives/negatives is explained in the following.

We have manually transcribed 1086 spoken utterances (in Portuguese), which were then labeled with the following tags, some depending on the answer given by EDGAR:

- 0: in-domain question incorrectly answered, although there was information in the knowledge sources (excluding Program D) to answer it;
- 1: out-of-domain question, incorrectly answered;
- 2: question correctly answered by Program D;
- 3: question correctly answered by using knowledge sources (excluding Program D);
- 4: in-domain question, incorrectly answered. There is no information in the knowledge source to answer it, but it should be;
- 5: multiple questions, partially answered;
- 6: multiple questions, unanswered;
- 7: question with implicit information (there, him, etc.), unanswered;
- 8: question which is not "ipsis verbis" in the knowledge source, but has a paraphrase there and was not correctly answered;
- 9: question with a single word (*garden*, *palace*), unanswered;
- 10: question that we do not want the system to answer (some were answered, some were not).

The previous tags were mapped into:

- true positives: questions marked with 2, 3 and 5;
- true negatives: questions marked with 0 and 10 (the ones that were not answered by the system);
- false positives: questions marked with 0 and 10 (the ones that were answered by the system);
- false negatives: questions marked with 4, 6, 7, 8 and 9.

Then, two experiments were conducted: in the first, the NLU module was applied to the manual transcriptions; in the second, directly to the output of the ASR. Table 1 shows the results.

NLU input = manual transcriptions		
Precision	Recall	F-measure
0.92	0.60	0.72
acNLU input = ASR		
Precision	Recall	F-measure
0.71	0.32	0.45

Table 1: NLU results

The ASR Word Error Rate (WER) is of 70%. However, we detect some problems in the way we were collecting the audio, and in more recent evaluations (by using 363 recent logs where previous problems were corrected), that error decreased to a WER of 52%, including speech from 111 children, 21 non native Portuguese speakers (thus, with a different pronunciation), 23 individuals not talking in Portuguese and 27 interactions where multiple speakers overlap. Here, we should refer the work presented in (Traum et al., 2012), where an evaluation of two virtual guides in a museum is presented. They also had to deal with speakers from different ages and with question off-topic, and report a ASR with 57% WER (however they majority of their user are children: 76%).

We are currently preparing a new corpus for evaluating the NLU module, however, the following results remain: in the best scenario, if transcription is perfect, the NLU module behaves as indicated in Table 1 (manual transcriptions).

5 Conclusions and Future Work

We have described a platform for developing ECAs with tutoring goals, that takes both speech and text as input and output, and introduced EDGAR, the butler of MONSERRATE, which was developed in that platform. Special attention was given to EDGAR's NLU module, which couples techniques that try to find distances between the user input and sentences in the existing knowledge sources, with a framework imported from the chatbots community (AIML plus Program D). EDGAR has been tested with real users for the last year and we are currently performing a detailed evaluation of it. There is much work to be done, including to be able to deal with language varieties, which is an important source of recognition errors. Moreover, the capacity of dealing with out-of-domain questions is still a hot research topic and one of our priorities in the near future. We have testified that people are delighted when EDGAR answers out-of-domain questions (*Do you like soccer?/I rather have a tea and read a good criminal book*) and we cannot forget that entertainment is also one of this Embodied Conversational Agent (ECA)'s goal.

Acknowledgments

This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013. Pedro Fialho, Sérgio Curto and Pedro Cláudio scholarships were supported under project FALACOMIGO (ProjectoVII em co-promoção, QREN n 13449).

References

N. O. Bernsen and L. Dybkjr. 2005. Meet hans christian andersen. In *In Proceedings of Sixth SIGdial Workshop on Discourse and Dialogue*, pages 237–241.

Tina Klwer. 2011. "i like your shirt" – dialogue acts for enabling social talk in conversational agents. In *Proceedings of the 11th International Conference on Intelligent Virtual Agents. International Conference on Intelligent Virtual Agents (IVA), 11th, September 17-19, Reykjavik, Iceland*. Springer.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2006. Building effective question answering characters. In *7th SIGdial Workshop on Discourse and Dialogue*, Sydney, Australia.

Hugo Meinedo, Diamantino Caseiro, João Neto, and Isabel Trancoso. 2003. Audimus.media: a broadcast news speech recognition system for the european portuguese language. In *Proceedings of the 6th international conference on Computational processing of the Portuguese language*, PROPOR'03, pages 9–17, Berlin, Heidelberg. Springer-Verlag.

H. Meinedo, A. Abad, T. Pellegrini, I. Trancoso, and J. P. Neto. 2010. The l2f broadcast news speech recognition system. In *Proceedings of Fala2010*, Vigo, Spain.

Ana Cristina Mendes, Rui Prada, and Luísa Coheur. 2009. Adapting a virtual agent to users' vocabulary and needs. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, IVA '09, pages 529–530, Berlin, Heidelberg. Springer-Verlag.

Sérgio Paulo, Luís C. Oliveira, Carlos Mendes, Luís Figueira, Renato Cassaca, Céu Viana, and Helena Moniz. 2008. Dixi — a generic text-to-speech system for european portuguese. In *Proceedings of the 8th international conference on Computational Processing of the Portuguese Language*, PROPOR '08, pages 91–100, Berlin, Heidelberg. Springer-Verlag.

Thies Pfeiffer, Christian Liguda, Ipke Wachsmuth, and Stefan Stein. 2011. Living with a virtual agent: Seven years with an embodied conversational agent at the heinz nixdorf museumsforum. In *Proceedings of the International Conference Re-Thinking Technology in Museums 2011 - Emerging Experiences*, pages 121 – 131. thinkk creative & the University of Limerick.

Susan Robinson, David Traum, Midhun Ittycheriah, and Joe Henderer. 2008. What would you ask a conversational agent? observations of human-agent dialogues in a museum setting. In *International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

David Traum, Priti Aggarwal, Ron Artstein, Susan Foutz, Jillian Gerten, Athanasios Katsamanis, Anton Leuski, Dan Noren, and William Swartout. 2012. Ada and grace: Direct interaction with museum visitors. In *The 12th International Conference on Intelligent Virtual Agents (IVA)*, Santa Cruz, CA, September.

Ulli Waltinger, Alexa Breuing, and Ipke Wachsmuth. 2011. Interfacing virtual agents with collaborative knowledge: Open domain question answering using wikipedia-based topic models. In *IJCAI*, pages 1896–1902.

Heiga Zen, Keiichiro Oura, Takashi Nose, Junichi Yamagishi, Shinji Sako, Tomoki Toda, Takashi Masuko, Alan W. Black, and Keiichi Tokuda. 2009. Recent development of the HMM-based speech synthesis system (HTS). In *Proc. 2009 Asia-Pacific Signal and Information Processing Association (APSIPA)*, Sapporo, Japan, October.