Luke, I am Your Father: Dealing with Out-of-Domain Requests by Using Movies Subtitles

David Ameixa[1], Luisa Coheur[1], Pedro Fialho[2], and Paulo Quaresma[2]

[1] Instituto Superior Técnico, Universidade de Lisboa/INESC-ID
[2] Universidade de Évora/INESC-ID
Rua Alves Redol n 9, 1000-029 Lisbon, Portugal
name.surname@inesc-id.pt

Abstract. Even when the role of a conversational agent is well known users persist in confronting them with Out-of-Domain input. This often results in inappropriate feedback, leaving the user unsatisfied. In this paper we explore the automatic creation/enrichment of conversational agents' knowledge bases by taking advantage of natural language interactions present in the Web, such as movies subtitles. Thus, we introduce Filipe, a chatbot that answers users' request by taking advantage of a corpus of turns obtained from movies subtitles (the Subtle corpus). Filipe is based on Say Something Smart, a tool responsible for indexing a corpus of turns and selecting the most appropriate answer, which we fully describe in this paper. Moreover, we show how this corpus of turns can help an existing conversational agent to answer Out-of-Domain interactions. A preliminary evaluation is also presented.

1 Introduction

The number of organisations providing virtual assistants is increasing. Examples of such assistants are Siri, from Apple, IKEA's Anna, or Monserrate's butler, Edgar Smith [5]. Yet, even when their roles are well known – for instance, answering questions about a specific domain or performing some pre determined task – users persist in confronting them with Out-of-Domain (OOD) input, that is, personal questions, requests about the weather, or about other topics unrelated with their tasks. Although it might be argued that such systems should only be focused in their own pre-defined functions, the fact is that people become more engaged with these applications if OOD requests are properly addressed [11]. Therefore, current approaches usually anticipate some OOD requests and handcraft answers for them. However, it should be clear that it is impossible to predict all the possible sentences that can be submitted to such agents.

An alternative solution to deal with OOD requests is to explore the (semi-)automatic creation/enrichment of the knowledge base of virtual assistants/chatbots by taking advantage of the vast amount of dialogues present in the web. Recently, Banchs and Li introduced *IRIS* [3], a chatbot that has in its knowledge base a corpus of interactions extracted from movie scripts (the *MovieDiC* corpus [2]). In this paper we take this idea one step further, and, instead of movie scripts, we propose the use of movie subtitles to

T. Bickmore et al. (Eds.): IVA 2014, LNAI 8637, pp. 13–21, 2014.

build a chatbot's knowledge base from scratch, and also to deal with the OOD requests of an existing conversational agent. Although less precise, Subtitles are easier to find and are available in almost every language; in addition, as large amounts of subtitles can be found, linguistic variability can be covered and redundancy can be taken into consideration (if a turn is repeatedly answered in the same way, that answer is probably a plausible answer to that turn). Therefore, in this paper we present Say Something Smart (SSS), a system that chooses an answer to a certain input, by taking into consideration a knowledge base of interactions – currently, the Subtle corpus [1], built from movies subtitles. We also show how this strategy can be applied to build a chatbot, Filipe (Figure 1)[1], and also to answer OOD requests posed by real users to Edgar Smith, the aforementioned Monserate's butler. Some preliminary results are also presented.

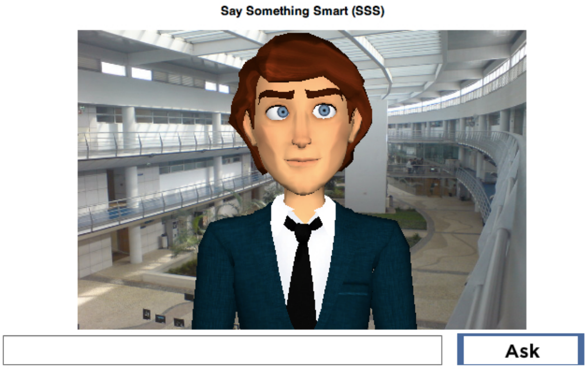


Fig. 1. Filipe, our chatbot based on SSS

This paper is organised as follows: in Section 2 we present some related work, in Section 3 we briefly describe the Subtle corpus, in Section 4 we detail SSS, and, in Section 5, we present a preliminary evaluation. Finally, in Section 6 we close the paper by providing some conclusions and pointing to some future work.

2 Related Work

Several conversational agents animate museums all over the world. Examples are: the 3D Hans Christian Andersen (HCA), which is capable of establishing multi-modal conversations about the namesake writer's life and tales [4]; Max, a virtual character employed as guide in the Heinz Nixdorf Museums Forum [12]; Sergeant Blackwell, installed in the Cooper-Hewitt National Design Museum in New York, and used by the U.S. Army Recruiting Command as a hi-tech attraction and information source [13]; the twins Ada and Grace, virtual guides in the Boston Museum of Science [15]. More recently, the virtual butler Edgar Smith [9,5] answers questions about Monserrate palace,

[1] Filipe can be tested in `http://www.l2f.inesc-id.pt/~pfialho/sss/`

in Sintra, Portugal, where he can be found. However, despite the sophisticated technologies behind all these systems, every single one reports the problem of having to deal with OOD requests.

In order to cope with this, these conversational agents follow different strategies: Edgar suggests questions when he is not able to understand an utterance, and starts talking about the palace if he does not understand the user repeatedly. A feature in his character also “excuses” some misunderstandings: as he is an old person, he does not have a very acute hearing; HCA changes topic when he is lost in the conversation, and also has an “excuse” for not answering some questions: the virtual HCA does not remember (yet) everything that the real HCA once knew; Max consults a Web-weather forecast for queries about this topic and uses Wikipedia to find answers to some factoid questions [16]. However, despite all these stratagems, actualisations of these agents knowledge bases, grounded in collected logs, still need to be performed on a regular basis.

The idea of taking advantage of existing human requests to feed dialogue systems emerges naturally from this context. Recently, the work presented in [3] describes a chat-oriented dialogue system, which has in its knowledge sources MovieDiC [2], a corpus extracted from movies scripts. In this paper we take this idea a little further and take advantage of movies subtitles to answer users’ requests. Contrary to movies scripts, subtitles exist in much larger quantities and for almost every language.

Considering the process of choosing the answer, we use SSS, which is based in Information Extractions techniques and also in edit distance metrics. The process behind SSS is somewhat similar to the one described in [6], where both a role-play (representing free-form human interactions) and a Wizard of Oz dialogue corpora (specific task turns) are used to find answers: our approach also works at the lexical level, not using any kind of dialogue act or semantic annotation. However, contrary to our approach, context is already taken into consideration.

Finally, another related systems that should be mentioned, although developed with other goal, is Say Anything [14] where the user and the computer take turns to write a story. Say Anything is based on a corpus of millions of stories extracted from weblogs and Lucene is also used by this system.

3 The Subtle Corpus

We follow [15] and envisage the use of knowledge bases constituted of turns. From now on we will call *interactions* to each pair of sentences (T, A), where A (the *answer*) corresponds to a response to T, from now on the *trigger*. The following are examples of interactions.

Example 1. (T1: You know, I didn’t catch your age. How old are you?, A1: 20)

Example 2. (T2: So how old are you?, A2: That’s none of your business)

The Subtle corpus is a collection of interactions, extracted from four different movies subtitles genres (Horror (H), Scifi (SF), Western (W) and Romance (R)), for Portuguese

Table 1. Number of available turns

#Interactions **Subtle – English**				
R	**SF**	**W**	**H**	**All**
1,392,569	625,233	333,776	1,000,902	3,352,480
#Interactions **Subtle – Portuguese**				
R	**SF**	**W**	**H**	**All**
627,368	477,521	129,081	696203	1,930,173

and English. Details on how this corpus was obtained from subtitles files can be found in [1]. Table 1 shows the number of available interactions for each genre.

It should be clear that as Subtle Interactions are obtained from subtitles' files based on the time elapsed between the dialogue lines, many turns in Subtle are not real dialogue pairs. In addition, some expected Interactions are not captured. This can be observed when Filipe is not able to answer *He told me YOU killed him.* with *No, I am your father*, a piece of dialogue from Star Wars[2]. This is due to the time elapsed between Darth Vader and Luke Skywalker lines that surpasses the time limit established to consider two turns as an interaction. Moreover, unexpected formats found in subtitles' files also led to false interactions. An example that illustrates this problem is: User: *Hasta la vista, babe*, SSS: *(CROWD CHEERING)*. Another example of unexpected information that can still be found in Subtle interactions is illustrated in the following, where the first turn shows a trigger with the name of the character that says it (PHILIP).

Example 3. (T3: PHILIP: How How are you?, A3: Fine.)

4 Say Something Smart

In the section we describe SSS. It takes as input the user request and returns an answer from the agent knowledge base. At the basis of this process there are several sentences comparisons. Thus, due to the large amount of interactions available, and because this selection needs to be done very fast (the user cannot wait long for an answer), a previous filtering step takes place in SSS before it selects an answer. In the following we detail these steps.

4.1 Indexing Subtle and Extracting Candidate Answers

We start by indexing the Subtle corpus through Lucene[3], a open-source, high-performance text search engine library, widely used by the Natural Language Processing/Information Extraction community. Then, given a user request, Lucene search engine is also used to retrieve a ranked set of interactions. Results returned by Lucene are based on an internal scoring algorithm[4] that takes into consideration the words present in the interactions and in the user request. Nevertheless, although the first interactions are usually

[2] Actually, *Luke, I'm your father* is a misquotation from Star Wars, parodied in other movies.

[3] `http://lucene.apache.org`

[4] `http://www.lucenetutorial.com/advanced-topics/scoring.html`

the most accurate, the fact is that several interactions in which the trigger is not semantically related with the user request, are also returned. For instance, given the user request *Do you have brothers?*, the following are examples of interactions returned by Lucene (in the first 20 positions).

Example 4. (T4: Brother! Do you've a cigarette?, A4: Take it.)

Example 5. (T5: Do you've any brothers?, A5: I'll manage the business)

Example 6. (T6: You don't have to go, brother., A6: I'm not your brother.)

Example 7. (T7: Brother, you don't have a clue., A7: What were you thinking about?)

Example 8. (T8: Didn't you have a brother in the war?, A8: Well, my brother Roy.)

In addition, SSS, as many Question/Answering systems (e.g. [7]), is based on the answers redundancy. That is, we assume that if an answer to a certain request is more frequent than others, it has a higher probability of being a plausible one. Thus, a "reasonable" number of interactions need to be returned by Lucene. Considering again time as a factor that needs to be taken into consideration, and after several preliminary experiments, we opt to ask Lucene for a maximum of 100 interactions, which guarantees redundancy but also allows SSS to obtain an answer to any question in less that one second (using an Intel Core i5-480M).

4.2 The Answer Selection Step

In order to choose an answer from the retrieved set of interactions, SSS performs two sequential tasks. As previously shown, many of the triggers from the interactions returned by Lucene might not be (semantically) related with the user request. Therefore, SSS starts by filtering the retrieved interactions, choosing only those where the triggers are similar to the user request, according to a given threshold. That is to say, all the retrieved interactions above the threshold are kept; all the others are discarded. Previous work [10] have shown us that a simple yet effective similarity measure is a combination between Jaccard similarity coefficient[5] and Overlap coefficient[6]. For Overlap we use bigrams with a minimum score of 0.4, and, for Jaccard, unigrams with minimum score of 0.7 (these values were empirically obtained). A weight-factor distributes the importance of both scores.

If no interactions is selected, a discarding answer such as *I'm sorry but I do not know how to answer your question* is given. Otherwise, SSS moves to a second step, where the answers of the remaining interactions are analysed. As previously said, as our approach is based on the answers redundancy, we check for the most frequent answer (see [8] for a review about answer selection in Question/Answering systems). Once again, we do not force exact matches, and answers are compared according with the previously mentioned similarity measure. If none of them is similar (above a threshold) to any of the remaining answers, a random answer is returned (which allows Filipe to sometimes provide different answers to the same input); otherwise, the most common answer, that is, the one that has the highest similar values concerning the other answers, is returned.

[5] http://en.wikipedia.org/wiki/Jaccard_index

[6] http://en.wikipedia.org/wiki/Overlap_coefficient

5 Preliminary Evaluation

5.1 Experimental Setup

We have at our disposal a corpus collected by the developers of Edgar Smith, representing requests posed by real people to Edgar. As expected, several OOD requests appear in these logs. From this corpus, the 58 questions unanswered by Edgar and marked as OOD were extracted and used in our first experiments, as described in the following.

5.2 Should We Take the Different Movies' Genres into Consideration?

Firstly, we wanted to understand how many OOD requests SSS was able to answer, and see if there was any significant difference in the capacity of the different movies genres to contribute with plausible answers. Therefore, we run SSS with the different partitions of the Subtle corpus, taking as input the previously mentioned 58 OOD requests. Results were labeled as:

- (Disc)arded, when SSS provides a discarding answer (e.g. User: *Why you talk so funny dude?* SSS: *I'm sorry, I'm not able to answer.*);
- OK when SSS returns a plausible answer (e.g. User: *Are you joking?* SSS: *Do I look like a joker?*);
- KO, when SSS supplies an inappropriate answer (e.g. User: *You have a big nose. SSS: I didn't say you kidnapped Megan*).

The attained results can be seen in Table 2.

Table 2. Evaluation of SSS answers

	Horror	**Romance**	**Western**	**Sci-fi**	**All Genres**
Disc	22	19	28	21	**16**
OK	20	21	17	23	**27**
KO	16	17	13	14	**15**

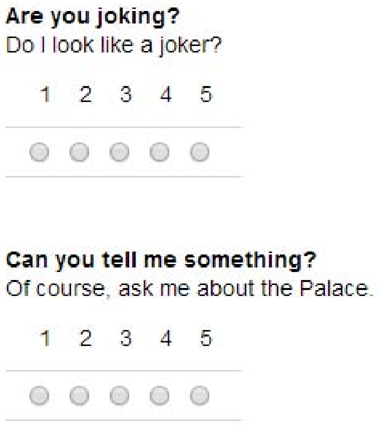
As expected, best results are obtained using the whole Subtle corpus (column **All Genres**). Answers from Westerns had the worst results and Sci-fi the best. Considering the all genres together, 72% of the requests are now answered (42 in 58), and, from these, about 65% are considered to be appropriate (27 in 58). The ones that were discarded pose no problem, as at the end the answer results in the same answer that Edgar would give: *I'm sorry, I'm not able to answer.*. Still, there are 26% of answers given that are not suitable to the users requests.

5.3 Answers Evaluation

From the 42 requests that did not return a discarded answer we selected 20 (from now on the test set), with the following criteria:

- Do not contain offensive language;
- From the 9 questions paraphrasing *How are you?* 3 where chosen randomly;
- Questions with only one word, like “here?” or “Where?”, were rejected;
- From the remaining requests, we randomly chose them.

Then, we built a questionnaire, based on the answers provided by SSS to those 20 requests, and also on hand-crafted plausible answers. Then we divided this questionnaire in two: each questionnaire had 10 questions of each type (generated by SSS and hand-crafted). To our 30 evaluators (adults with different ages and backgrounds, not necessarily working in computer science) we told that these requests were posed to Edgar and that the answers were its response. Evaluators should give them a 1-5 score according to how satisfied they were (being 5 the best score). A snippet of the questionnaire can be found in Figure 2.



Are you joking?
Do I look like a joker?

1 2 3 4 5
○ ○ ○ ○ ○

Can you tell me something?
Of course, ask me about the Palace.

1 2 3 4 5
○ ○ ○ ○ ○

Fig. 2. Snippet from the questionnaire

To test the evaluators concordance we used the Cronbach α measure. The inter-rater agreement score was high, for both the SSS and the manual answers (0,80 and 0,84, respectively). People preferred the manual answers, as expected: SSS obtained a score of 2,9 $\pm 1, 37$ and manual answers of 4,3 $\pm 0, 84$.

The variation of results attained from SSS also show that some of the provided questions are very good, and others are very bad. Five of our questions had an average rate of less than 1.9 points; in the other hand the remaining answers had an average rate of almost 4.0. Analysing the 5 answers with a very low rate, we can see that they are not related with the topic of the question or they not fit in the present context, like when is said *You have a big nose*, and the answer is *I didn't say you kidnapped Megan*. The same happens when is given the utterance *I said Hello!* and the answer is *I say shut up, you bilge rat, before I use you as an anchor.* and with the question *Are you a donkey?* that is answered with *Ok. Do one thing. Tell me the story of your journey to heaven once again. Only once*. Some answers, although "funny", are definitely not suitable to the environment where, for instance, Edgar is integrated (an elegant palace).

Thus, despite some problems related with the corpus itself, SSS strategy needs to be improved so that answers related with a very specific context are avoided.

Nevertheless, this strategy can easily provide answers to many (not so obvious) OOD requests. For instance, if Filipe is told *I like your hair* or *Are you a mutant?*, he will provide very reasonably answers.

6 Conclusions and Future Work

We have presented an approach to deal with OOD requests that takes advantage on movies subtitles. Thus, we have built SSS a tool that indexes the interactions from its knowledge base (the Subtle corpus, obtained from movies subtitles) and, given an user request, chooses an answer. With Subtle and SSS we have created Filipe, a virtual agent, implemented over SSS. Moreover, we have used this system to answer OOD requests asked by real users to Edgar Smith, a virtual butler operating in Monserrate palace.

Although much work still needs to be done regarding the subtle corpus and SSS, the fact is that several questions that the butler was unable to deal with can now be successfully answered; moreover, some answers given by Filipe are extremely interesting, if we consider that his developers did not have to hand-craft them. Nevertheless, it should be clear that such approach needs to be improved before being applied to a formal agent such as Edgar (answers need to be customise to be adequate to an old butler, slang needs to be eliminated, etc.). Thus, future work includes the refinement of Subtle, so that badly formed turns are discarded, and the organisation of the corpus, so that paraphrases are detected, as well as very specific answers. Moreover, normalisation of the corpus is one of our targets, and we intent to use a named entity recogniser to generalise turns involving names entities. With respect to SSS, a main concern is to extended its way of selecting an appropriate answer, as many important variables, such as context (as done in [6]) and the agent personality should be taken into account.

Acknowledgments. This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013.

References

1. Ameixa, D., Coheur, L.: From subtitles to human interactions: introducing the subtle corpus. Tech. rep., INESC-ID (November 2014)
2. Banchs, R.E.: Movie-dic: a movie dialogue corpus for research and development. In: Proceedings of the 50th Annual Meeting of the ACL, pp. 203–207. Association for Computational Linguistics, Jeju Island (2012)

3. Banchs, R.E., Li, H.: Iris: a chat-oriented dialogue system based on the vector space model. In: ACL (System Demonstrations), pp. 37–42 (2012)
4. Bernsen, N.O., Dybkjær, L.: Meet hans christian anderson. In: Proceedings of the Sixth SIGdial Workshop on Discourse and Dialogue, pp. 237–241 (2005)
5. Fialho, P., Coheur, L., dos Santos Lopes Curto, S., Cládio, P.M.A.: Ângela Costa, Abad, A., Meinedo, H., Trancoso, I.: Meet edgar, a tutoring agent at monserrate. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL (August 2013)
6. Gandhe, S., Traum, D.: First steps towards dialogue modelling from an un-annotated human-human corpus. In: 5th Workshop on Knowledge and Reasoning in Practical Dialogue Systems, Hyderabad, India (January 2007)
7. Mendes, A.C., Coheur, L., Silva, J., Rodrigues, H.: Just.ask - a multi-pronged approach to question answering. International Journal on Artificial Intelligence Tools 22(1) (2013)
8. Mendes, A.C., Coheur, L.: When the answer comes into question in question-answering: survey and open issues. Natural Language Engineering 19(1), 1–32 (2013)
9. Moreira, C., Mendes, A.C., Coheur, L., Martins, B.: Towards the rapid development of a natural language understanding module. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 309–315. Springer, Heidelberg (2011)
10. Mota, P., Coheur, L., Curto, S., Fialho, P.: Natural language understanding: From laboratory predictions to real interactions. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2012. LNCS, vol. 7499, pp. 640–647. Springer, Heidelberg (2012)
11. Patel, R., Leuski, A., Traum, D.R.: Dealing with out of domain questions in virtual characters. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 121–131. Springer, Heidelberg (2006)
12. Pfeiffer, T., Liguda, C., Wachsmuth, I., Stein, S.: Living with a virtual agent: Seven years with an embodied conversational agent at the heinz nixdorf museumsforum (2011)
13. Robinson, S., Traum, D.R., Ittycheriah, M., Henderer, J.: What would you ask a conversational agent? observations of human-agent dialogues in a museum setting. In: LREC. European Language Resources Association (2008)
14. Swanson, R., Gordon, A.S.: Say anything: A massively collaborative open domain story writing companion. In: First International Conference on Interactive Digital Storytelling, Erfurt, Germany (November 2008)
15. Traum, D., Aggarwal, P., Artstein, R., Foutz, S., Gerten, J., Katsamanis, A., Leuski, A., Noren, D., Swartout, W.: Ada and grace: Direct interaction with museum visitors. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 245–251. Springer, Heidelberg (2012)
16. Waltinger, U., Breuing, A., Wachsmuth, I.: Interfacing virtual agents with collaborative knowledge: Open domain question answering using wikipedia-based topic models. In: Walsh, T. (ed.) Proceedings of the 22nd IJCAI, IJCAI 2011, Barcelona, Spain, July 16-22, pp. 1896–1902. IJCAI/AAAI (2011)