From subtitles to human interactions: introducing the `SubTle` Corpus

David Ameixa, Luísa Coheur
IST/INESC-ID
Rua Alves Redol, 9
1000-029 Lisboa
luisa.coheur@inesc-id.pt

October 17, 2013

Abstract

Domain-oriented dialogue systems are often faced with users that try to cross the limits of their knowledge, by unawareness of its domain limitations or simply to test their capacity. These interactions are considered to be Out-Of-Domain and several strategies can be found in the literature to deal with some specific situations. Since if a certain input appears once, it has a non-zero probability of being entered later, the idea of taking advantage of real human interactions to feed these dialogue systems emerges, thus, naturally. In this paper, we introduce the `SubTle` Corpus, a corpus of Interaction-Response pairs extracted from subtitles files, created to help dialogue systems to deal with Out-of-Domain interactions.

1 Introduction

Dialogue systems can be divided into task-oriented and non-task-oriented systems, being the latter also called domain-oriented systems [1]. While task-oriented systems objectives are achieved by completing a set of subtasks [2, 3], domain-oriented systems do not have these task constrains to guide the users. Although some constrains can be imposed by making explicit the system domain of expertise (for instance, Edgar Smith [4], the butler of Monserrate Palace, presents itself by saying that he "only" answers questions about the palace), the fact is that people tend to challenge these systems with interactions that cannot be handled by their knowledge base – the so called Out-of-Domain (OOD) interactions.

Some systems are prepared to answer to specific OOD interactions (for instance, a general Question/Answering system was integrated in Max [5], a virtual character employed as guide in the Heinz Nixdorf Museums Forum), or by simply expressing that the system is unable to handle it [6]. However, addressing OOD in a proper way, specially in small talk situations, can help establishing a closer relation with the user; at the same time it will give the system more credibility [7].

In this context, taking human interactions as an information source that can be used to answer such questions seems to be a natural way to deal with OODs. Some works are already starting to approach this challenge. A recent example is the chatbot presented in [8], IRIS, that uses Movie-DiC [9], a corpus extracted from movies' scripts, to find answers to its users' interactions.

In this work, we take advantage of an easier to find information source (although not so informative, as the movies' scripts): English and Portuguese subtitles. Based on these, we create a corpus, the SubTle Corpus, that we use to deal with the OOD interactions presented to Edgar Smith. Although not being the focus of this paper, some initial tests were already made, and results shows that 46.5% of the OODs submitted to Edgar are now successfully answered.

This paper is organised as follows: in Section 2 we describe the task of collecting subtitles; in Section 3 we explain the process of creating Interaction-Response (I-R) pairs based on the previous attained corpus, and, in Section 4 we present some statistics of the corpus and a preliminary evaluation. Finally, in Section 5 we present some conclusions and future work.

2 Collecting subtitles

In this section we explain the process of gathering the subtitles that were used to create the SubTle corpus.

The first step was to search in IMDB[1] for lists with movies' names. We focused on four genres: Horror, Scifi, Western and Romance. Then, we extracted, along with each movie, its IMDBid, that is, the number that univocally identifies it, which is used by the cataloging system of IMDB, and also widely used by the community (in order to obtain these numbers, we used a free API[2]). Then, we used an open-source tool called vroksub[3] to connect to the database of opensubtitles[4] and, using an API based on a simple protocol called XML-RPC[5], searched and downloaded the required subtitles. The best subtitles were chosen according with its format – SubRip (srt) – and its rating (calculated according to the score users assign to that subtitle).

However, since we were limited to download only 200 subtitles a day, we contacted the administrator of opensubtitles and he gently sent us the remaining subtitles we needed, consisting of about 2GB of subtitles in Portuguese and English, in the four aforementioned genres. However, the previous described process can be adopted to get subtitles from any other site that allows a search using the movies' identification and whose administrators do not provide any help.

3 From Subtitles to Interactions-Response pairs

As previously mentioned, the subtitles used to build the SubTle corpus are in srt format. This format consists of a set of "slots" where each slot comprises four parts, all in text:

1. A number indicating the position of the slot in the sequence.
2. The time that the slot should appear on the screen, and then disappear.
3. The content of the subtitle.
4. A blank line indicating the start of a new slot.

After an analysis of several subtitles, we concluded that there are two possibilities regarding the

[1] http://www.imdb.com

[2] http://www.omdbapi.com/

[3] https://code.google.com/p/vroksub/

[4] http://opensubtitles.org

[5] http://trac.opensubtitles.org/projects/opensubtitles/wiki/XMLRPC

content of each slot: they are due to the same character or belongs to two different characters. Also, it was necessary to analyze slots not individually but sequentially. Therefore, given a sequence of slots, we separate the interventions of each character, based on the following:

- If the first slot begins with a capital letter or hyphen and ends with a punctuation mark indicating the end of a sentence (.!?), the second slot too and both are separated by a time smaller than X seconds (this corresponds to the time difference between the end of a slot and the beginning of another – in this paper we take X = 1 second), then they are considered as belonging to the same dialogue and form a I-R pair. It is important to note that this value of 1 second was chosen by an empirical analysis and can be easily changed.

- If the first slot begins with capital letter, ends with characters that do not indicate the end of the sentence, like "..." or with a lower case, and the second slot begins with a sentence with the same kind of characters, these slots are considered as corresponding to the same character. In this case, no I-R pair is generated.

4 The `SubTle` Corpus

The `SubTle` corpus aggregates different groups of I-R pairs, extracted from four different movies subtitles genres (Horror, Scifi, Western and Romance), for Portuguese and English. Details can be seen in the next Table.

In a preliminary evaluation, we selected 15 slots from a random subtitle file, from each one of the existent genres. For each one of them, we manually built a reference, containing the I-R pairs that should be created. The attained recall was of 97% and precision of 71%. This means that, using this strategy, 29% of the generated corpus is not well formed, but, in the other hand, 97% of the I-R pairs that were supposed to be created, are being created and only 3% are wrongly discarded.

`SubTle` – Portuguese				
	Corpus (Genders)			
	Romance	**Sci-fi**	**Western**	**Horror**
#original subtitle files	1121	1069	277	1703
#subtitles files discarded	41	29	6	42
#I-R pairs attained	627368	477521	129081	696203
Average I-R pairs per subtitle file	580	459	476	419
`SubTle` – English				
	Corpus (Genders)			
	Romance	**Sci-fi**	**Western**	**Horror**
#original subtitle files	2127	1310	616	2131
#subtitles files discarded	36	33	7	36
#I-R pairs	1392569	625233	333776	1000902
Average I-R pairs per subtitle file	477	496	533	496

5 Conclusions and Future Work

We presented the `SubTle` corpus, built from subtitles, with the goal of being used to deal with OOD interactions within a domain-oriented dialogue system. In order to create the `SubTle` corpus, we proposed an algorithm that takes subtitles in the SubRip format and extract I-R pairs from it. In the final paper, we will show several examples to properly illustrate this process.

A preliminary evaluation shows a recall (97%) higher than precision (71%), considering 1 second as a delimiter for an interaction.

Some problems regarding this corpus are: a) existence of errors (many of the subtitles are made by amateurs); b) absence of a standard to write subtitles, leading to the impossibility of creating an accurate algorithm to extract the I-R pairs.

Future challenges will be to discard badly formed I-R pairs in the corpus, and take context into account, as the specific context of some dialogues, lead to I-R pairs that hardly fit in any chatbot alike future dialog. To be able to properly organize this corpus is one of our main concerns.

References

[1] Dybkjær, L., Bernsen, N.O., Minker, W.: Evaluation and usability of multimodal spoken language dialogue systems. In: IN: SPEECH COMMUNICATION. (2004) 33–54

[2] an Joseph Polifroni, S.S.: Dialogue management in the Mercury Flight Reservation System. In: Proceedings of the ANLP/NAACL 2000 Workshop on Conversational Systems, Seattle, Association for Computational Linguistics (May 2000) 11–16

[3] Allen, J.F.: The trains project: A case study in building a conversational planning agent. Journal of Experimental and Theoretical AI JETAI **7** (1995) 7–48

[4] Fialho, P., Coheur, L., dos Santos Lopes Curto, S., Cláúdio, P.M.A., Costa, A., Abad, A., Meinedo, H., Trancoso, I.: Meet edgar, a tutoring agent at monserrate. In: ACL. Proceedings of the 51st Annual Meeting of the Association f (August 2013)

[5] Waltinger, U., Breuing, A., Wachsmuth, I.: Interfacing virtual agents with collaborative knowledge: Open domain question answering using wikipedia-based topic models. In: IJCAI. (2011) 1896–1902

[6] Patel, R., Leuski, A., Traum, D.: Dealing with out of domain questions in virtual characters. In: IVA 2006. LNCS (LNAI, Springer (2006)

[7] Bickmore, T., Cassell, J.: ”how about this weather?” - social dialogue with embodied conversational agents (August 09 2000)

[8] Banchs, R.E., Li, H.: Iris: a chat-oriented dialogue system based on the vector space model. In: ACL (System Demonstrations). (2012) 37–42

[9] Banchs, R.E.: Movie-dic: a movie dialogue corpus for research and development. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Jeju Island, Korea, Association for Computational Linguistics (July 2012) 203–207