A Spanish Corpus for Talking to the Elderly

Raquel Justo, Leila Ben Letaifa, Javier Mikel Olaso, Asier López-Zorrilla, Mikel Develasco, Alain Vázquez, M. Inés Torres

Abstract

In this work, a Spanish corpus that was developed, within the EMPATHIC project [1] framework, is presented. It was designed for building a dialogue system capable of talking to elderly people and promoting healthy habits, through a coaching model. The corpus, that comprises audio, video an text channels, was acquired by using a Wizard of Oz strategy. It was annotated in terms of different labels according to the different models that are needed in a dialogue system, including an emotion based annotation that will be used to generate empathetic system reactions. The annotation at different levels along with the employed procedure are described and analysed.

[1] http://www.empathic-project.eu/

0.1 Introduction

Although the use of conversational systems in our daily life seemed to be science fiction not much time ago. Nowadays they are pretty integrated in our homes (Alexa speaker by Amazon), jobs (Cortana or Siri to manage our agenda) or even in our leisure (Siri or Samsung's Bixby for smartphones). They are becoming useful in more and more different domains ranging from game environments to educational contexts. Some of them can pass the Turing test (e.g., Eugene Goostman [1]). Thus, we can say that the way in which people interact with computers is shifting to the use of natural language.

There are many different systems in the literature built for different purposes and that make use of different technologies [27, 2, 23, 9]. However, one of the most extended categorization of conversational systems is the one that distinguishes among "chatbots" and "dialogue systems" [11, 20, 14]. Although the frontiers among those categories are not always clear, focusing on the differences related to the goal, chatbots are aimed at generating appropriate, relevant, meaningful and human-like utterances and there is not an specific goal to be achieved during the interaction like in the case of dialogue systems. Dialogue systems are often developed for a specific domain, whereas simulated conversational systems [chatbots] are aimed at open domain conversation [15].

In this work we deal with a dialogue system developed within the EMPATHIC project [25, 26, 13] framework. The goal of this project is to design and validate new interaction paradigms for personalized Virtual Coaches to promote healthy and independent aging. Thus, a dialogue system that can talk to the elderly, understand them, empathize with them and promote healthy habits is being developed. This kind of dialogue systems need different modules like automatic speech recognizer, natural language understanding module, dialog manager, natural language generator, etc. Moreover, a module that tries to detect the emotion of the speaker is also being developed in order to provide a system response that can be empathetic with regard to the user emotional status. The methodologies employed to develop these modules are mainly based on machine learning and data driven approaches. When using these approaches, data are needed to be able to train robust models. Moreover, the data have to be closely related to the specific task, environment, channel, etc. Thus, it is very difficult to get valuable resources when specific tasks, like the one presented in this framework are considered. Furthermore the lack of resources is even more noticeable when we consider other languages (apart from English) like Spanish.

The main contribution of this paper is the description of a Spanish corpus devoted to train different models that will be employed in a dialogue system that tries to talk to the elderly people and promotes healthy habits being aware of the affective component. The corpus was annotated in terms of different labels that will be used by the different modules. The annotation procedures, that will be described in the following sections, were selected to allow the Dialog Manager to understand the user in terms of the coaching strategies and goals to be developed and agreed with the user, which is a challenging and novel approach. Section 0.2 provides a description of the dialogues that comprise the corpus and the way in which they were acquired. In Section 0.3 the annotation procedure developed to build the modules related to dialogue generation are described (natural language understanding, dialog manager and natural language generation). Then in Section 0.4 the annotation carried out to detect emotions in different channels (audio, video and text) is detailed.

0.2 Dialogues in the EMPATHIC framework

In order to develop a dialogue system, like the one described above, a data acquisition procedure has to be designed first. In this process we used a Wizard of Oz (WoZ) platform [21, 22] for the acquisition of the database. The WoZ constitutes a prototyping method that uses a human operator (the so-called wizard) to simulate non- or only partly- existing system functions. It was used to make users think that they are interacting with a real automatic dialogue system. In this way, the data acquisition procedure considers human-machine conversations that were carried out in an environment as most realistic as possible.

The dialogues in the EMPATHIC project are leaded according to a coaching model, a GROW coaching model in this case, that tries to get the desired goals related to healthy habits. A GROW coaching dialogue consists on four main phases: Goals or objectives, Reality, Options and Will or action plan. During the first phase, the dialogue aims at establishing explicit objectives that the user wants to achieve, e.g. reduce the amount of salt. During the next phase, taking into account the user's personal context, the dialogue tries to detect potential obstacles that prevent fulfilling the previously established objectives. For the next phase, the goal is to analyse the options the user has in order to face the obstacles and achieve the objectives. In the last phase, the dialogue tries to specify an action plan for the user to carry out in order to advance towards the objectives. The final goal for the EMPATHIC virtual coach is to deal with four different domains: leisure, nutrition, physical activity and social and family relationships [19]. However, in the initial phase described in this paper, not all the scenarios were used; two scenarios were integrated in this platform. A first introductory scenario, which in turn was used to strengthen the user in the interaction with the platform. And a second one to simulate a GROW session on nutrition. These scenarios were designed using the documentation provided by a professional coach. Although different acquisition procedures were carried out in the project for different languages: Spanish, French and Norwegian, in this work we focus on the Spanish dataset.

Making use of the aforementioned WoZ platform, 79 native Spanish users selected among the target population (healthy elderly above 65) interacted with the system. The majority of them participated in the two predefined scenarios, but in some cases, due to different reasons, only one of these sessions was carried out. Thus, 142 dialogues were collected. These include around 4,500 user turns and the same amount of machine turns.

The acquired conversational sessions between elderly people and the simulated virtual coach were recorded in order to have an audio-visual database. Each session takes about 10 minutes so the total recordings correspond to about 23 hours of video. The audio part represents about 30% of the database.

0.3 Resources for building the dialogue

Once the acquisition procedure was finished the data were annotated in order to build the different models involved in the conversational process.

0.3.1 Speech to Text Annotation

One of the first annotation needed for training robust models to be used in a dialogue system is the transcription of the speech. This is essential for the Automatic Speech Recognizer for instance. Thus after the acquisition procedure, the dialogues were manually transcribed. The vocabulary size resulted to be 5,543 for the user turns and 2,941 for the virtual coach's turns. As for the running words, the corpus contains 72,350 in the user turns and 30,389 in the he virtual coach's turns.

The transcriptions of the acquired dialogues were further annotated in order to facilitate the modeling of the dialogues. The following two sections explain how the turns of both the users and the virtual coach were labeled. The two annotation tasks were carried out by 9 annotators, who were instructed about the structure of the labels, the GROW coaching model, and about the context of the project. Each annotator labeled roughly the same number of dialogues.

0.3.2 Semantic Annotation

The taxonomy of the labels used to represent each of the users' utterances was designed so as to be usable for the dialogue agent that is being deployed in the EMPATHIC project. Several works have addressed the question of defining dialogue act taxonomies [5, 24]. Among them, the DIT++ taxonomy [4] and the more recent ISO 24617-2 standard [17, 6], which is intended to be a development of the

Topic	Intent	Entities
sport & leisure - travelling	*generic - agreement*	*actions*
sport & leisure - hobbies - type	*GROW - habit - present*	*quantities*
nutrition - regularity - ordered	*generic - opinion - positive*	*places*
sport & leisure - motivation	*generic - disagreement*	*amount of time*
nutrition - quantity	*generic - greetings*	*frequencies*
sport & leisure - music	*GROW - plan*	*hobbies*

Table 1: Most frequent topic, intent and entity labels in the corpus.

previous one, can be considered the general methodological framework of the taxonomy defined in this section. It is a dialogue-act taxonomy aimed to represent the user utterances in a particular human-machine communication framework, which develops a coaching model aimed at keeping a healthy and independent life of elderly. Thus, the taxonomy allows the Dialog Manager to understand the user in terms of the coaching strategies and goals to be developed and agreed with the user, which is a challenging and novel approach. To fulfill its needs, we employed three different types of labels: the topic, the intent and the name entities. The topic label identifies the general context of utterance, such as nutrition, leisure or family; and also some subtopics. The intent label determines the communicative intention of the user, e.g. greetings, agreement, opinion and so on. Additionally, it also includes information about which stage of the GROW model the user is talking about. Finally, the name entities are tuples containing small segments of the utterance and their category. They can be very useful for understanding the user but also for enriching the natural language generator. We have included, for example, people's names, places, and books.

The topic and intent labels are hierarchical, i.e., each utterance is labeled with multiple tags that can be ordered from more general to more specific. To make the annotation more consistent, each turn was split into several subsentences if there were more than one topic or intents in that turn. In total, 56 different labels were used for the topic representation, 34 for the communicative intent and 22 types of entities were identified. The complete list of labels is provided in detail in [16]. Since it is too large, Table 1 shows the most frequent labels for the topic and intent, and the most frequent entities.

0.3.3 Dialogue Act annotation

Dialogue Act (DA) annotation is the equivalent task to the semantic annotation for the turns of the virtual coach. In this case, the outputs of the coach are labelled considering five criteria: DA, polarity, gender of the user and coach and possible appearance of entities in the responses of the coach. Such annotation is highly related to the Natural Language Generation (NLG), one of the modules included in the dialogue system developed in the EMPATHIC project. The NLG is in charge of generating the responses of the virtual coach to the users through a unit of information which contains a set of labels. The inverse process is made in the annotation: one set of labels is assigned to each turn of the virtual coach contained in the data.

The data was extracted from two different sources: the WoZ sessions and a set of handmade dialogues prepared by a professional coach. In both cases, only the turns of the coach were relevant to build this part of the data. Indeed, each turn can be split in different utterances, where an utterance is considered each element which can be labelled with a different DA. In total, the number of utterances is 8173 where 5985 are from the real session with users and 2188 from the handcrafted conversations.

All these utterances were labelled in terms of the five aforementioned criteria. The DA, which is built for one principal label and sublabel in the case of EMPATHIC, describes the communicative function and the semantic of the coach's sentences. There are 10 different values for the principal label and more than 100 for the sublabel. However, the DAs do not allow all the possible combinations, as each label only can be joined with a reduced group of sublabels. The polarity defines the emotional state of the coach, which can be selected between positive and neutral. The possible values for the genders are male, female and not identifiable, since what is annotated is if the gender of the two participants can be known through the coach turn alone (without any context). Finally, the detection of entities followed the same procedure carried out in the semantic annotation.

In the DAs, we identified three different blocks with the following distribution: the GROW block (19.6%), the Introduction one (24.6%) and General one (55.8%). The first block contains eighth of the ten principal labels. These labels are the eight typical questions used in the GROW model. The other blocks, each one only contains one principal label. The Introduction label is used to annotate usual turns in a first session with the user (ask for the name, self-introduction, information of the project, ...). Finally, the General one is used to label all the expression which can be part of any conversation (thanking, greetings, agreement, ...). In terms of the polarity, the positive utterances (63.0%) were almost two times the neutral ones (37.0%). For both user and coach gender, they were not identifiable in almost the 99% of the data. Finally, the most frequent entities in the data were actions, dates and food.

0.4 Resources for empathizing with the elderly

Within the EMPATHIC project framework, the idea of empathizing was very important. Thus, we wanted not only to understand what elderly is requesting to the system, but also to know their emotional status when interacting with it. Therefore, an annotation in terms of emotion was carried out by Spanish native annotators. The representation of emotional status is not straightforward and different models can be used according to Affective Computing literature [7, 8, 18, 3]. In this work we employed both a categorical model and a three-dimensional VAD (Valence, Arousal and Dominance) model in order to be able to compare both criteria.

Both data modalities, audio and video, were considered. In order to avoid interference between modalities, only audio (i.e. no images) was provided to the speech annotators and only video (without sound) was used by the video annotators.

In this section, we describe and analyse each modality annotation. For more information about the annotation procedure, refer to [10] and [12].

Finally, at the same time as the semantic annotation was carried out, the polarity of the transcribed utterances was also labeled by the same annotators.

0.4.1 Audio Annotation

Only the audio part of the conversations between the virtual coach and elder people (which duration is about 7 hours) is concerned by the audio annotation process. A manual labeling procedure from scratch was carried out by three native people. The perceived emotion was labelled in terms of categorical labels and the three-dimensional VAD model labels. The labels assigned to the dimensional VAD model were:

- Valence: Positive, Neither positive nor negative, Negative
- Arousal: Excited, Slightly excited, Neutral
- Dominance: Dominant, Neither dominant nor intimidated, Defensive

The categorical labels were Calm, Sad, Happy, Puzzled and Tense. For each emotion label, the number of segments labeled by each annotator is reported in Table 2. ”Calm” is the most frequent label. ”Happy” and ”Puzzled” are less present but ”Sad” and ”Tense” are quite absent.

Dealing with the duration of emotion labels, “Calm” occurs in 94% of the audio database size which correspond to more than 6 hours. ”Happy” and ”Puzzled” labels are present in only 4% of the database with respective duration’s of 9 and 8 minutes.The negative emotions “Sad” and ”Tense” have a total duration lower than one second. This could indicate that the dialog system is user friendly.

	Calm	***Sad***	***Happy***	***Puzzled***	***Tense***
Annotation 1	7017	17	260	347	12
Annotation 2	7794	19	291	297	24
Annotation 3	7655	21	244	360	20

Table 2: Audio annotated segments

0.4.2 Video Annotation

For the video annotation, all the database is involved and the recordings were labeled by two native people. Six video emotion categories were selected: Sad, Annoyed/Angry, Surprised, Happy/Amused, Pensive and Other. The label Other is assigned to segments containing different emotions that the sub-mentioned ones or including simultaneous emotions.The non annotated parts are considered neutral. For each emotion label, the number of segments labeled by each annotator is reported in Table 3.

With respectively 0, 1 and 3 occurrences, ”Sad”, ”Annoyed” and ”Other” are almost absent. ”Pensive” and ”Neutral” represent the most frequent labels. Indeed, as more than 14 hours are not labeled, the content of the database is mainly neutral. The participants are annotated ”Pensive” within a duration of about 2 hours. Finally they are sometimes happy or amused (during 5-10 minutes).

0.4.3 Text Annotation

Emotions were not only labeled from audio and video (sections 0.4.1 and 0.4.2) but also from text, that is from the manual transcriptions achieved in 0.3.1. It was carried out along with the semantic annotation (0.3.2) providing n emotional annotation for each transcribed utterance.

Although the audio and video has richer annotations, the text annotation includes a very significant one related to polarity labels on a scale of three values: negative, neutral and positive. This might be very useful to be combined with the audio annotation in terms of the VAD model. Specifically, the combination of Valence (audio) and Polarity (text) labels we can get the same annotation for

	Sad	***Annoyed***	***Surprised***	***Happy***	***Pensive***	***Other***	***Neutral***
Annotation1	0	0	12	234	2032	0	2278
Annotation2	0	1	44	151	2059	3	2258

Table 3: Video annotated segments

different channels.

Looking at the annotated set it can be concluded that Neutral is the most common polarity, representing the 66.24% of the corpus, then a positive behaviour can be analyzed, with a 27.21% of the corpus, and finally, negative polarity is almost absent (with 6.55% of occurrences).

0.5 Concluding Remarks

In this work a Spanish corpus devoted to the development of a dialogue system, oriented to promoting healthy habits among elderly is presented. The corpus was annotated in terms of different labels in order to obtain robust models for generating coherent system responses according to a coaching model. Moreover, an emotion-based annotation is also provided in order to detect emotional status of the speakers and provide a response adapted to it. The procedure carried to obtain the annotations along with the obtained results is described.

Bibliography

[1] Eugene: Turing test sucess marks milestone in computing history (2014). University of Reading Press Releases

[2] Reward estimation for dialogue policy optimisation. Computer Speech Language **51**, 24 – 43 (2018). DOI https://doi.org/10.1016/j.csl.2018.02.003

[3] Bradley, M.M., Lang, P.J.: Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry **25**(1), 49 – 59 (1994). DOI https://doi.org/10.1016/0005-7916(94)90063-9. URL http://www.sciencedirect.com/science/article/pii/0005791694900639

[4] Bunt, H.: The dit++ taxonomy for functional dialogue markup. In: AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts, pp. 13–24 (2009)

[5] Bunt, H., Alexandersson, J., Choe, J.W., Fang, A.C., Hasida, K., Petukhova, V., Popescu-Belis, A., Traum, D.R.: ISO 24617-2: A semantically-based standard for dialogue annotation. In: LREC, pp. 430–437 (2012)

[6] Bunt, H., Petukhova, V., Malchanau, A., Wijnhoven, K., Fang, A.: The dialogbank. In: N.C.C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (2016)

[7] Calvo, R.A., D'Mello, S.: Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on Affective Computing **1**(1), 18–37 (2010). DOI 10.1109/T-AFFC.2010.1

[8] Calvo, R.A., Mac Kim, S.: Emotions in text: Dimensional and categorical models. Computational Intelligence **29**(3), 527–543 (2013). DOI 10.1111/j.1467-8640.2012.00456.x

[9] Gao, J., Galley, M., Li, L.: Neural Approaches to Conversational AI: Question Answering, Task-oriented Dialogues and Social Chatbots (2019)

[10] Justo, R., Letaifa, L.B., Palmero, C., Gonzalez-Fraile, E., Johansen, A., Vazquez, A., Cordasco, G., Schlogl, S., Fernandez-Ruanova, B., Silva, M., Escalera, S., Velasco, M.D., Tenorio-Laranga, J., Esposito, A., Kornes, M., Torres, M.: Analysis of the interaction between elderly people and a simulated virtual coach. Ambient Intelligence and Humanized Computing (In Press)

[11] Klüwer, T.: From Chatbots to Dialog Systems Chapter. Conversational Agents and Natural Language Interaction: Techniques and Effective Practices, chap. 8, pp. 1–22. IGI Global Publishing (2011)

[12] L. Ben Letaifa, T.M.I., Raquel, J.: Adding dimensional features for emotion recognition on speech. In: International Conference on advanced technologies for signal and image processing, pp. 109–114. Tunisia (2020)

[13] López Zorrilla, A., Velasco Vázquez, M.d., Irastorza, J., Olaso Fernández, J.M., Justo Blanco, R., Torres Barañano, M.I.: Empathic: Empathic, expressive, advanced virtual coach to improve independent healthy-life-years of the elderly. Procesamiento del Lenguaje Natural (2018)

[14] Masche, J., Le, N.T.: A review of technologies for conversational systems. In: Advanced Computational Methods for Knowledge Engineering, pp. 212–225. Springer International Publishing, Cham (2018). DOI 10.1007/978-3-319-61911-8_19

[15] Mctear, M.: Spoken dialogue technology - toward the conversational user interface. (2004)

[16] Montenegro, C., López Zorrilla, A., Mikel Olaso, J., Santana, R., Justo, R., Lozano, J.A., Torres, M.I.: A dialogue-act taxonomy for a virtual coach designed to improve the life of elderly. Multimodal Technologies and Interaction **3**(3), 52 (2019)

[17] Petukhova, V., Bunt, H.: The coding and annotation of multimodal dialogue acts. In: LREC, pp. 430–437 (2012)

[18] Russell, J.: Core affect and the psychological construction of emotion. Psychological Review **110**, 145–172 (2003). DOI 10.1037/0033-295X.110.1.145

[19] Sayas, S.: Dialogues on Leisure and Free Time, Dialogues on Physical Exercise, Dialogues on Nutrition. Technical Report DP1, DP2, DP3, Empathic Project; Internal Documents: Tampere, Finland (2018)

[20] Scerri, D., Dingli, A.: Dialog systems and their inputs. In: C. Stephanidis (ed.) HCI International 2013 - Posters' Extended Abstracts, pp. 601–605. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)

[21] Schlögl, S.: Wizard of Oz Prototyping Framework. [Online]. Available: https://github.com/stephanschloegl/WebWOZ

[22] Schlögl, S., Milhorat, P., Chollet, G., Boudy, J.: Designing language technology applications: A wizard of Oz driven prototyping framework. In: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 85–88. Association for Computational Linguistics, Gothenburg, Sweden (2014). DOI 10.3115/v1/E14-2022

[23] Serban, I., Sordoni, A., Bengio, Y., Courville, A.C., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: AAAI (2015)

[24] Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C.V., Meteer, M.: Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational linguistics **26**(3), 339–373 (2000)

[25] Torres, M.I., Olaso, J.M., Glackin, N., Justo, R., Chollet, G.: A spoken dialogue system for the empathic virtual coach. In: L.F. D'Haro, R.E. Banchs, H. Li (eds.) 9th International Workshop on Spoken Dialogue System Technology, pp. 259–265. Singapore (2019)

[26] Torres, M.I., Olaso, J.M., Montenegro, C., Santana, R., Vázquez, A., Justo, R., Lozano, J.A., Schlögl, S., Chollet, G., Dugan, N., Irvine, M., Glackin, N., Pickard, C., Esposito, A., Cordasco, G., Troncone, A., Petrovska-Delacretaz, D., Mtibaa, A., Hmani, M.A., Korsnes, M.S., Martinussen, L.J., Escalera, S., Cantariño, C.P., Deroo, O., Gordeeva, O., Tenorio-Laranga, J., Gonzalez-Fraile, E., Fernandez-Ruanova, B., Gonzalez-Pinto, A.: The empathic project: Mid-term achievements. In: Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, PETRA '19, pp. 629–638. ACM, New York, NY, USA (2019). DOI 10.1145/3316782.3322764

[27] Williams, J.D., Asadi, K., Zweig, G.: Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In: ACL (2017)