Universitat Politècnica de Catalunya

Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona

A Generative Dialogue System for Reminiscence Therapy

Mariona Carós Roca

Advisors: Xavier Giró-i-Nieto, Petia Radeva

A thesis submitted in fulfillment of the requirements for the Master in Telecommunications Engineering

Barcelona, September 2019

Abstract

With people living longer than ever, the number of cases with neurodegenerative diseases such as Alzheimer's or cognitive impairment increases steadily. In Spain it affects more than 1.2 million patients and it is estimated that in 2050 more than 100 million people will be affected. While there are not effective treatments for this terminal disease, therapies such as reminiscence, that stimulate memories of the patient's past are recommended, as they encourage the communication and produce mental and emotional benefits on the patient. Currently, reminiscence therapy takes place in hospitals or residences, where the therapists are located. Since people that receive this therapy are old and may have mobility difficulties, we present an AI solution to guide older adults through reminiscence sessions by using their laptop or smartphone.

Our solution consists in a generative dialogue system composed of two deep learning architectures to recognize image and text content. An Encoder-Decoder with Attention is trained to generate questions from photos provided by the user, which is composed of a pretrained Convolution Neural Network to encode the picture, and a Long Short-Term Memory to decode the image features and generate the question. The second architecture is a sequence-to-sequence model that provides feedback to engage the user in the conversation.

Thanks to the experiments, we realise that we obtain the best performance by training the dialogue model with Persona-Dataset and fine-tuning it with Cornell Movie-Dialogues dataset. Finally, we integrate Telegram as the interface for the user to interact with Elisabot, our trained conversational agent.

Resum

El nombre de casos amb malalties neurodegeneratives com l'Alzheimer o el deteriorament cognitiu augmenta constantment. A Espanya afecta més d'1,2 milions de pacients i es calcula que el 2050 es veuran més de 100 milions de casos. Si bé no hi ha tractaments eficaços per a aquesta malaltia, es recomanen teràpies com la reminiscència, que estimulen els records del pacient i fomenten la comunicació. Actualment, la teràpia de reminiscència es fa en hospitals o residències, que és on es troben els terapeutes. Com que les persones que reben aquesta teràpia són grans i poden tenir dificultats de mobilitat, presentem una solució basada en Intel·ligència Artificial per realitzar sessions de reminiscència mitjançant l'ús del portàtil o del telèfon mòbil.

La nostra solució consisteix en un sistema de diàleg generatiu format per dues arquitectures d'aprenentatge profund que reconeixen el contingut d'imatges i de text. Per una banda, un Codificador-Descodificador amb *Attention* per generar preguntes basades en el contingut de les fotografies, proporcionades pels usuaris, format per una xarxa neuronal convolucional (CNN) que codifica les imatges i una LSTM que genera les preguntes paraula a paraula. La segona arquitectura consisteix en un model *sequence-to-sequence* que genera comentaris a les respostes dels usuaris per enriquir la conversa.

Després de realitzar diversos experiments, veiem que obtenim el millor comportament entrenant el model de conversa amb les dades de *Persona-chat* i ajustant el model amb *fine-tune* de *Cornell Movie-Dialogues*. Finalment, integrem *Telegram* com a interfície perquè l'usuari interactuï amb la Elisabot.

Resumen

El número de casos con enfermedades neurodegenerativas como el Alzheimer o el deterioro cognitivo aumenta de manera constante. En España afecta a más de 1,2 millones de pacientes y se estima que en 2050 se verán afectados más de 100 millones de personas. Si bien no existen tratamientos efectivos para esta enfermedad terminal, se recomiendan terapias como la reminiscencia, que estimulan los recuerdos del pasado y fomentan la comunicación del paciente. Actualmente, la terapia de reminiscencia se realiza en hospitales o residencias, donde se encuentran los terapeutas. Dado que las personas que reciben esta terapia son mayores y pueden tener dificultades de movilidad, presentamos una solución basada en inteligencia artificial para guiar a los usuarios a través de sesiones de reminiscencia utilizando su portátil o teléfono inteligente.

Nuestra solución consiste en un sistema de diálogo generativo compuesto por dos arquitecturas de aprendizaje profundo para reconocer el contenido de imagen y texto. Por un lado, un Codificador-Descodificador con *Attention* para generar preguntas basadas en el contenido de las fotografías proporcionadas por los usuarios formado por una red neuronal convolucional (CNN) que codifica las imágenes y una LSTM que genera las preguntes palabra a palabra. La segunda arquitectura consiste en un modelo *sequence-to-sequence* que genera comentarios a las respuestas de los usuarios para enriquecer la conversación..

Después de realizar varios experimentos, vemos que obtenemos el mejor comportamiento entrenando el modelo con los datos de *Persona-chat* y ajustando el modelo con *fine-tune* de *Cornell Movie-Dialogue*. Finalmente, integramos *Telegram* como interfaz porque el usuario interactue con nuestro agente Elisabot.

Acknowledgements

First of all, I want to thank my advisor Xavier Giró-i-Nieto for guiding me during the project development, updating me with the newest technologies and link me to professionals in the field. I would also like to thank professor Petia Radeva for her support whenever she could and the director of the clinical research group Maite Garolera who gave me her advice from the medical point of view and made possible to test the proposed work with patients of *Consorci Sanitari de Terrassa*.

This project emerged from *Vodafone Campus Lab* and deserves particular credit for this. Its main goal is the granting of scholarships for the development of masters or doctoral projects related with digital transformation. The proposal of this thesis was presented to give a solution to the question: "How could we create a solution for people affected by cognitive impairment or beginning of Alzheimer, together with their families, to store their best memories so that they feel again the happiest moments of their lives and improve their quality of life?". I am very grateful to Vodafone for giving me the opportunity to develop this work and specially to Mari Satur Torre, Marcos Bou and Estíbaliz Ochoa for the talks in the meetings we had during the development of the project and the contacts of experts in reminiscence therapy.

There are many people who has contributed to this work through his professional advice, such as the director of CEAFA Jesús, machine learning scientists Elia Bruni from Universitat Pompeu Fabra, Agata Lapedriza from the Universitat Oberta de Catalunya, Marta Ruiz from the Universitat Politecnica de Catalunya, and the responsible of projects for old people in Cruz Roja Joaquín. Thank you to all of you.

Revision history and approval record

Revision	Date	Purpose
0	13/06/2019	Document creation
1		Document revision
2		Document revision
3		Document approbation

DOCUMENT DISTRIBUTION LIST

Name	e-mail
Mariona Carós Roca	mariona.caros@estudiant.upc.edu
Xavier Giró i Nieto	xavier.giro@upc.edu
Mari Satur Torre	mari-satur.torre@vodafone.com

Written by:		Reviewed and approved by:		Reviewed and approved by:	
Date		**Date**		**Date**	
Name	Mariona Carós Roca	**Name**	Xavier Giró-i-Nieto	**Name**	
Position	Project Author	**Position**	Project Supervisor	**Position**	Project Supervisor

Contents

List of Figures

List of Abbreviations

AI	**A**rtifitial **I**ntelligence
API	**A**pplication **P**rogramming **I**nterface
NN	**N**eural **N**etwork
CNN	**C**onvolutional **N**eural **N**etwork
RNN	**R**ecurrent **N**eural **N**etwork
LSTM	**L**ong **S**hort-**T**erm **M**emory
ResNet	**Res**idual **Net**work
GPU	**G**raphics **P**rocessing **U**nit
RT	**R**eminiscence **T**herapy
VQG	**V**isual **Q**uestion **G**eneration

Chapter 1

Introduction

1.1 Motivation

With people living longer than ever, the number of cases with neurodegenerative diseases such as Alzheimer's increases steadily [9][13][19]. Research focused on identifying treatments to slow down the evolution of neurodegenerative diseases is a very active pursuit, but it has been only successful in terms of developing therapies that eases the symptoms without addressing the cause [1][27]. Besides, not all the people with dementia might have accessibility to the therapy, as it requires to pay a specialized therapist that might be expensive and move to the specific hospital or residence where the therapy takes place. We believe that artificial intelligence can contribute in innovative systems to give accessibility and offer new solutions to the patients needs, as well as help relatives and caregivers to understand the illness of their family member or patient and monitor the progress of the dementia.

Cognitive impairment or Alzheimer's disease affects more than 1.2 million patients in Spain and up to 6 million people indirectly, as family members and carers, and it is estimated that in 2050 more than 100 million people will be affected [19]. Therapies such as Reminiscence, that stimulate memories of the patient's past, has well documented benefits on social, mental and emotional well-being [29][10], making it a very desirable practice, especially for older adults. Reminiscence Therapy (RT) in particular involves the discussion of events and past experiences using tangible prompts such as photographs or music to evoke memories and stimulate conversation [39].

With this aim, we explore multimodal deep learning architectures to be used to develop an intuitive, easy to use, and robust dialogue system for people at early stages of Alzheimer's disease or cognitive impairment. The main goal of this work is the development of a system able to generate questions from images and maintain a conversation with an old adult. We divide this goal in the following tasks:

- Explore the techniques used for dialogue and visual question generation (VQG).
- Find open-source data bases to properly train our models.
- Develop an agent that supports effective conversation about the life of a person.
- Integrate both models (VQG and chatbot) in an interface to generate the Reminiscence Therapy.
- Evaluate the therapy with two patients diagnosed of cognitive impairment.

This thesis is structured as follows. A brief introduction to artificial intelligence and dialogue systems and an overview of the related work in this field is given in Chapter 2. The implemented system is described in Chapter 3 and evaluated in a series of experiments in Chapter 4. Chapter 5 gives an estimation of the costs associated to the development of the project. Discussion and future research directions are provided in Chapter 6. Finally, appendices include the evaluation survey that was given to users.

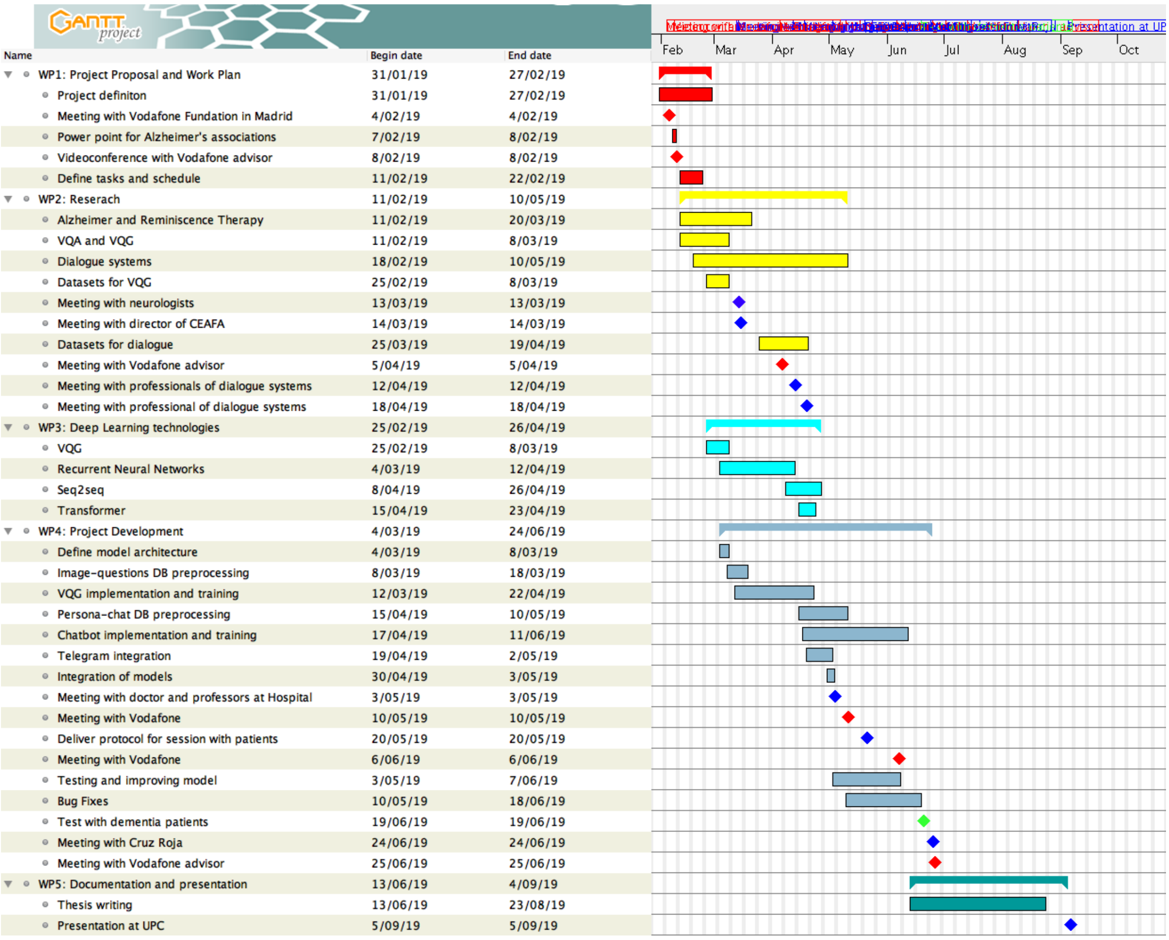
1.2 Hardware and Software Resources

This project was developed using the NVIDIA GPU of Image Processing Group at the Universitat Politècnica de Catalunya (UPC).

The algorithm was implemented with Pytorch[1], using CUDA and cuDNN for fast GPU primitives. This framework was chosen because it is a Python-based deep learning library which is open source, builds applications on top of dynamic graphs which can be played with on runtime. Code is publicly available at GitHub[2]. We borrowed code from the Pytorch chatbot tutorial[3] and Pytorch Tutorial Image Captioning[4].

1.3 Work Plan

This project has followed the attached work plan.



Name	Begin date	End date
WP1: Project Proposal and Work Plan	31/01/19	27/02/19
Project definiton	31/01/19	27/02/19
Meeting with Vodafone Fundation in Madrid	4/02/19	4/02/19
Power point for Alzheimer's associations	7/02/19	8/02/19
Videoconference with Vodafone advisor	8/02/19	8/02/19
Define tasks and schedule	11/02/19	22/02/19
WP2: Reserach	11/02/19	10/05/19
Alzheimer and Reminiscence Therapy	11/02/19	20/03/19
VQA and VQG	11/02/19	8/03/19
Dialogue systems	18/02/19	10/05/19
Datasets for VQG	25/02/19	8/03/19
Meeting with neurologists	13/03/19	13/03/19
Meeting with director of CEAFA	14/03/19	14/03/19
Datasets for dialogue	25/03/19	19/04/19
Meeting with Vodafone advisor	5/04/19	5/04/19
Meeting with professionals of dialogue systems	12/04/19	12/04/19
Meeting with professional of dialogue systems	18/04/19	18/04/19
WP3: Deep Learning technologies	25/02/19	26/04/19
VQG	25/02/19	8/03/19
Recurrent Neural Networks	4/03/19	12/04/19
Seq2seq	8/04/19	26/04/19
Transformer	15/04/19	23/04/19
WP4: Project Development	4/03/19	24/06/19
Define model architecture	4/03/19	8/03/19
Image-questions DB preprocessing	8/03/19	18/03/19
VQG implementation and training	12/03/19	22/04/19
Persona-chat DB preprocessing	15/04/19	10/05/19
Chatbot implementation and training	17/04/19	11/06/19
Telegram integration	19/04/19	2/05/19
Integration of models	30/04/19	3/05/19
Meeting with doctor and professors at Hospital	3/05/19	3/05/19
Meeting with Vodafone	10/05/19	10/05/19
Deliver protocol for session with patients	20/05/19	20/05/19
Meeting with Vodafone	6/06/19	6/06/19
Testing and improving model	3/05/19	7/06/19
Bug Fixes	10/05/19	18/06/19
Test with dementia patients	19/06/19	19/06/19
Meeting with Cruz Roja	24/06/19	24/06/19
Meeting with Vodafone advisor	25/06/19	25/06/19
WP5: Documentation and presentation	13/06/19	4/09/19
Thesis writing	13/06/19	23/08/19
Presentation at UPC	5/09/19	5/09/19

Figure 1.1: Gantt Diagram

[1] https://www.pytorch.org/

[2] https://github.com/marionacaros/generative-dialogue-system-for-reminiscence-telegram

[3] https://pytorch.org/tutorials/beginner/chatbot_tutorial.html

[4] https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning

Chapter 2

State of the art

Artificial Intelligence arises from the aim of creating machines that simulate human intelligence. Software able to do anything that human beings can do, such as understand an article, recognize an image, speak, or even drive a car, to help us in our every day life and improve it.

It is known that machines are faster and more precise than human beings solving mathematical problems or even playing chess or Go [36], but when it comes to other tasks that we do without thinking such as recognizing a familiar face, it becomes more complicated for them to solve. Deep Learning presents a method to solve these more intuitive problems by allowing machines to learn from experience.

Landing these ideas from a mathematical point of view, the goal of a neural network is to approximate some function $y = f(x;\theta)$ by learning the value of the parameters θ that result in the best function approximation. A neural network is composed by layers, as the number of layers increase the function becomes more complex and the capacity of the model increases as well. But, not only the algorithm is important in deep learning, data plays a crucial roll. As the amount of training data increases, the amount of complexity required to achieve a good performance reduces. The network is provided with labels, training and validation data about what has to be the output, but the behavior of the layers is not defined, so the network has to keep tuning the parameters to optimize a cost function, that measures how well those predictions correspond with expected outputs, until it converges to a minimum and the network is able to generate what it has been asked to.

There are different neural network architectures depending on the type of data and the task required to model. We will review some of the literature involved in the process of building a conversational agent for reminiscence using neural networks. In this work we focus on the main supervised neural networks: Feedforward NN, Recurrent NN (RNN) [38][23] and Convolutional NN (CNN) [16].

2.1 Visual Question Generation

The first task we want to carry out is being able to generate questions given an image. In the area of language and vision, one of the most widely studied areas is image captioning, which involves generating a description given an image [35][41], which not differs greatly from our task. It has been shown [5][6] that a pretrained CNN on a large dataset paired with a language model such as a RNN with word embeddings promise a good performance on image captioning.

In Visual Question Generation (VQG), the system is tasked with asking an engaging question when shown an image. The first work to tackle this topic was [20] which focus on questions that are interesting for a person to answer, trying to get information that could not be known by only looking at the image. They provide three datasets with a total of 75,000 questions that we use in this work and explain in section 3.2, which range from object to event-centric images.

2.2 Conversational Agents

One of the goals in the field of AI is to build computer systems that can have human-like conversations with users. However the origin of chatbots goes back to 1966 with the creation of ELIZA [37] by Joseph Weizenbaum at MIT. Its implementation consisted in pattern matching and substitution methodology. Recently, data driven approaches have drawn significant attention. Existing work along this line includes retrieval-based methods [12][34][40] and generation-based methods[24][25]. In this work we focus on generative models, where sequence-to-sequence algorithm that uses RNNs to encode and decode inputs into responses is a current best practice.

Building an open-domain conversational agent is a challenging problem. As addressed in [42] and [7], the lack of a consistent personality and lack of long-term memory which produces some meaningless responses in these models are still unresolved problems.

Other works have proposed conversational agents for a variety of uses such as palliative care [32] or daily assistance. An example of a virtual assistant is 'Billie' reported in [15] which is a virtual agent that uses facial expression and head movement for a more natural behavior and is focused on managing user's calendar, or 'Mary' [31] that assists the users by organizing their tasks offering reminders and guidance with household activities. Both of the works perform well on its specific tasks, but report verbal miss-understandings and difficulties to maintain a casual conversation.

An important aspect to obtain good results when training a generative model is to chose a dataset that contains the type of data you want to generate. We looked for datasets of reminiscence therapies, but we did not find any, so we focused on datasets with casual conversations and visually grounded dialogues, as these are the type of dialogues used in reminiscence. After a comprehensive search, we concluded there is a lack of good open source available datasets for casual dialogues. Most of dialogue's datasets are goal-oriented, such as restaurant booking [3] or comprehension of text [22].

Some of the available datasets for open-dialogue generation are:

- Cornell-Movie Dialogue corpus [4] which contains a large collection of fictional conversations extracted from raw movie scripts, the problem of this dataset is that some of the conversations are oriented to movie topics like killing someone or find the murderer.
- DailyDialog [17] includes conversations about our daily life, but dialogues are mostly limited to domains appropriate for use as a language learning tool such as asking for directions.
- Reddit dataset [11] is composed by casual conversations from Reddit posts.
- Persona-Chat [42] is composed of dialogues between two persons with specific profiles that are trying to know each other.

There are works that propose a dataset of dialogues grounded in images where a conversation is conducted based on a given photo. It is the case of Image-Chat and PhotoBook datasets.

- Image-Chat [26] is a dataset that contains images and short dialogues about these images between two speakers, each of one playing the role of a given personality.

- PhotoBook [8] is a large-scale collection of visually-grounded dialogues task-oriented to describe the picture.

Other works use different techniques, such as [43] that uses a graph attention mechanism with large scale commonsense knowledge to facilitate language understanding or [11] that uses Reinforcement Learning from a fixed batch of human interaction data to generate open-domain dialogue.

Chapter 3

Methodology

3.1 Framework

The proposed system aims at engaging the user. We named it *Elisabot* and its main interest is to know about user's life events. Before starting the conversation, the user have to introduce photos that should contain significant moments or important people in his/her life. The system randomly chooses one of this pictures and analyses the content. Then, Elisabot shows the selected picture and starts the conversation by asking a question about the picture. The user should give an answer even though he does not know it and Elisabot makes a relevant comment on it. The cycle starts again by asking another relevant question about the image and the flow is repeated for 4 to 6 times until the picture is changed. The Figure 3.1 summarizes the behavior of our algorithm.

Elisabot is composed of two models: The model in charge of asking questions about the image, which we will refer to it as VQG model in the rest of the document, and the Chatbot model which tries to make the dialogue more engaging by giving feedback to the user's answers.

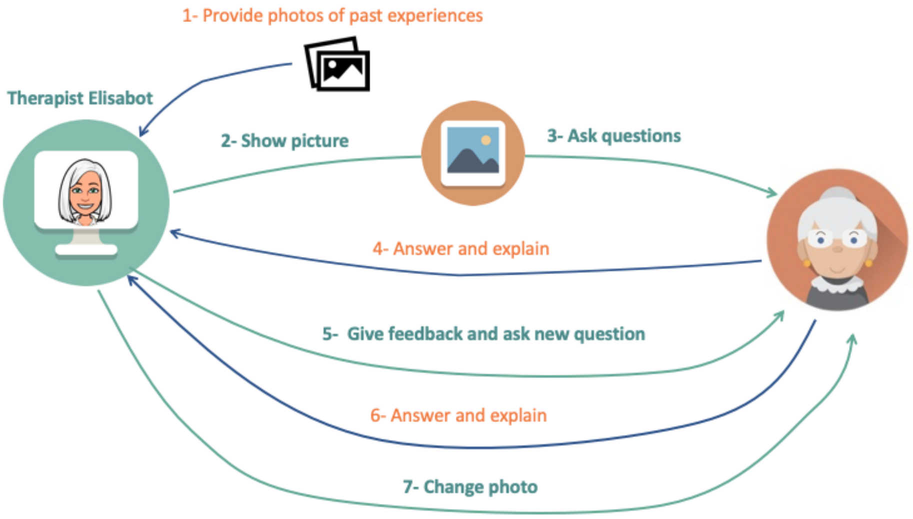


Figure 3.1: Scheme of the algorithm behavior

3.2 Characteristics of datasets

In order to achieve a good question generator, we looked for a dataset with natural questions that could engage someone to start a conversation. Computer vision questions that could be answered by looking at only the image, were outside the scope of this task.

The dataset we used to train the VQG model is composed of three datasets: MS COCO, Bing and Filckr, which contain a wide range of visual elements and situations. Each source contains 5,000 images with 5 questions per image, adding a total of 15,000 images with 75,000 questions. Coco dataset [18] includes images of complex everyday scenes containing common objects in their natural context, but it is limited in terms of the concepts it covers. In the Figure 3.2 we can see the 40 most frequent words and realise that the dataset is significantly pet biased. Bing dataset contains more event related questions and has a wider range of questions longitudes (between 3 and 20 words), while Flickr questions are shorter (less than 6 words) and the images appear to be more family related.

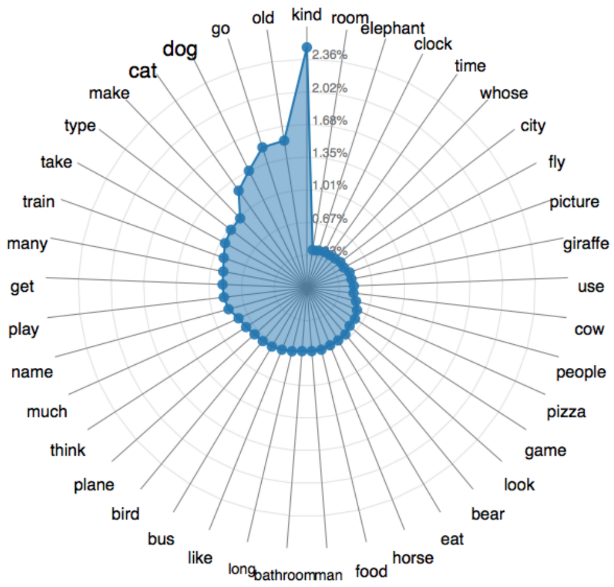


Figure 3.2: Frequency graph of top 40 words in COCO dataset, reference [20]

The pretrained Resnet used in our VQG model is trained on ImageNet which is a popular dataset of millions of labeled high-resolution images belonging to 22k categories.

The chatbot was trained with Persona-chat and Cornell-Movie Dialogues dataset. Persona-chat dataset includes dialogues between people that is getting to know each other and sentences have a maximum of 15 words, making it easier to learn for machines. In total it has 162,064 utterances (sequences of words) over 10,907 dialogues. Cornell-Movie dataset contains a collection of fictional conversations extracted from raw movie scripts. In total it has 304,713 utterances.

The following lines are an example of Persona-chat:

```
- hi !  i work as a gourmet cook
+ i don t like carrots .  i throw them away
- really .  but i can sing pitch perfect
+ i also cook and i ride my bike to work
- great !  i had won an award for spelling bee
+ okay but i was published in new yorker once
- you better not make any spelling mistakes
```

Here, there is a sample of Cornell-Movie dataset:

```
++++++ KAT ++++++ When you were gone last year -- where were you?
++++++ PATRICK ++++++ Busy
++++++ KAT ++++++ Were you in jail?
++++++ PATRICK ++++++ Maybe.
++++++ KAT ++++++ No, you weren’t
++++++ PATRICK ++++++ Then why’d you ask?
++++++ KAT ++++++ Why’d you lie?
++++++ KAT ++++++ I should do this.
++++++ PATRICK ++++++ Do what?
++++++ KAT ++++++ This.
++++++ PATRICK ++++++ Start a band?
++++++ KAT ++++++ My father wouldn’t approve of that that
++++++ PATRICK ++++++ You don’t strike me as the type that would ask permission.
++++++ KAT ++++++ Oh, so now you think you know me?
++++++ PATRICK ++++++ I’m gettin’ there
++++++ PATRICK ++++++ So what ’ s up with your dad?  He a pain in the ass?
++++++ KAT ++++++ He just wants me to be someone I’m not.
```

3.3 Preprocessing

To use the datasets in our model and obtain a good performance, we first must process the data so the model gets the useful information. We do it in the following way.

In the case of images, we first download them from datasets URLs and we find that some of them look like Figure 3.3. Some of them are not available anymore, so we filter them by using the pixel values as well as the size, because all the wrong pictures look the same and are smaller than they should. Once we have the valid images (around 10.000), we resize them all to 256x256 for uniformity, and normalize the pixel values by the mean and standard deviation of ImageNet which are: `mean = [0.485, 0.456, 0.406]` and `std = [0.229, 0.224, 0.225]`.

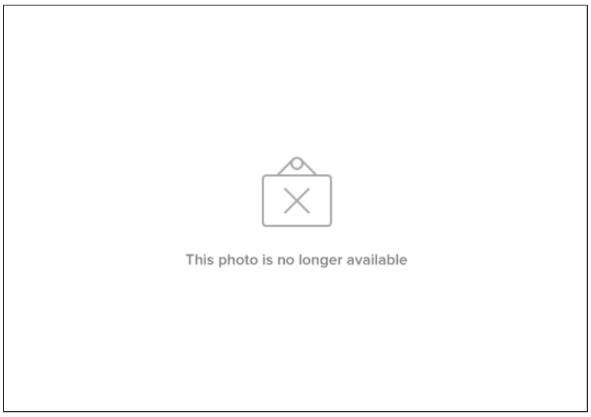


Figure 3.3: Picture not available.

In the case of text, we create two dictionaries of words, each one including words from the datasets to be used in each of the models (VQG and Chatbot). Each unique word is mapped to an index, which is going to be its identifier in the model. We count how many times a word appears in the dataset and we trim the words that appear less than 3 times because decreasing the feature space soften the difficulty of the function that the model learns to approximate.

Our models only accepts ASCII encoding, so we convert the Unicode strings to ASCII. Next, we convert all letters to lowercase and trim all non-letter characters except for basic punctuation, so the model is able to understand all the input information.

We split our datasets in batches of 32 training samples to be trained at same time to speed up training. Since we use fixed size batches in our model and sentences have different lengths, we need to add padding at those with fewer words. To accommodate sequences of different lengths in same batch we make our batches (`max_length, batch_size`), where sentences shorter than a maximum length are zero padded after an <`end`> token. We save the length of each sequence to be able to decode the batch. Finally, we need to index our batch along time so that the model learns to generate the words in order for all sequences in the batch. Therefore, we transpose the batch to return a time step across all sentences in the batch.

In the case of questions, after several experiments, we decide to filter questions larger than 6 words as we want simple questions easy to understand for our target users. Besides, filtering long sentences makes it easier for the model to converge. For the Chatbot model we trim pairs of sentences that have a sequence larger than 12 words. We reduce our Persona-chat training dataset from 137.101 to 39.753 sentence pairs and the Cornell dataset from 221.282 to 84.836 sentence pairs.

3.4 VQG Network

The algorithm behind VQG is based on the model of *Show, Attend and Tell* [41] which consists in a Encoder-Decoder architecture with Attention. The Encoder takes as input one of the given photos $\boldsymbol{I}$ from the user and learns its information using a CNN. CNNs have been widely studied for image tasks and are the state-of-the art for object recognition. The CNN provides the image's learned features to the Decoder which generates the question $\boldsymbol{y}$ word by word by using an attention mechanism with a Long Short-Term Memory (LSTM), which is a type of RNN architecture. It is natural to use a RNN to generate sequences as it is designed to learn temporal knowledge. The model is trained to maximize the likelihood $p(y|I)$ of producing a target sequence of words:

$$\mathbf{y} = \{y_1, ..., y_c\}, y_i \in \mathbb{R}^K \tag{3.1}$$

where K is the size of the vocabulary and C is the length of the caption.

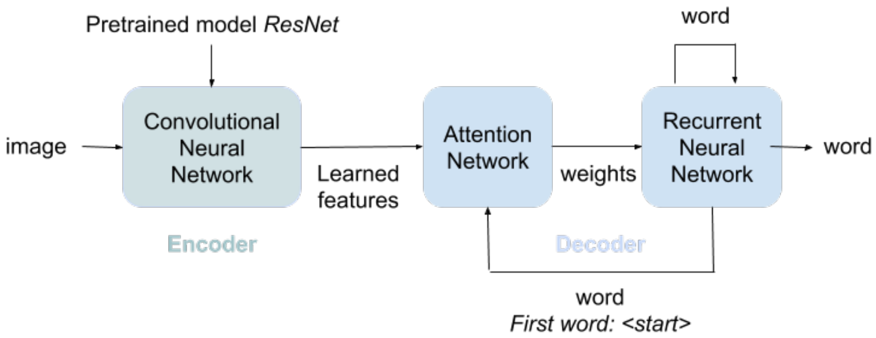


Figure 3.4: VQG architecture.

Since there are already CNNs trained on large datasets to represent images with an outstanding performance, we make use of transfer learning to integrate a pre-trained model into our algorithm. In particular, we use a *101-layered Residual Network* trained on *ImageNet*. We discard the last 2 layers, since these layers classify the image into categories and we only need to extract its features. The output of the encoder is a matrix of 14x14 with 2048 channels, which is the learned representation of the original image.

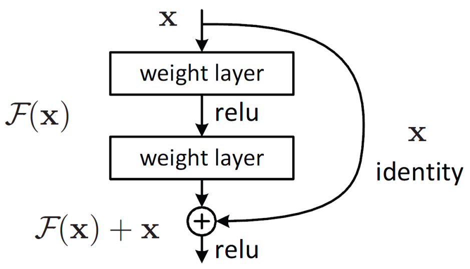


Figure 3.5: The Residual Network (ResNet) introduces shortcut connections to fit the input from the previous layer to the next layer without any modification of the input. Skip connections enable to avoid vanishing and exploding gradients because even if there is vanishing gradient for the weight layers, we always still have the identity $\mathtt{x}$ to transfer back to earlier layers, thus it enables to have a deeper network with better performance.

As we said before, the decoder is composed by an Attention Network and LSTM. We specifically use a LSTM because it deals with exploding and vanishing gradient problems that can be encountered when training traditional RNN by incorporating gates to regulate the flow of information. The LSTM produces a word at every time step conditioned on the weights from the Attention Network containing the visual information, the previous hidden state and the previously generated words.

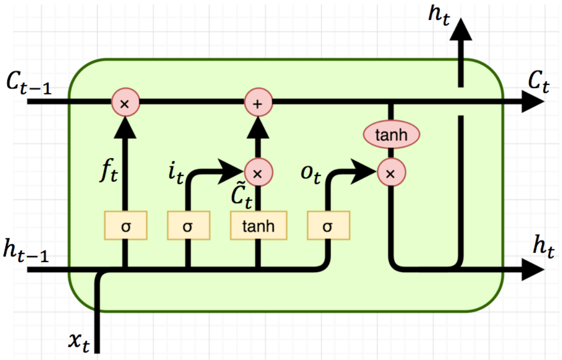


Figure 3.6: The core of the LSTM is a memory cell C_t encoding knowledge at every time step of observed inputs. The behavior of this cell is controlled by the input gate i_t, the output gate o_t and the forget gate f_t.

The Attention Network is composed by linear layers and a couple of activations. By using the previous generated word from the LSTM, it generates the weights that indicate where to focus within the image to generate the following word. At the first time step the LSTM will not have generated any word yet, so we use <start> as the first input to the Attention Network.

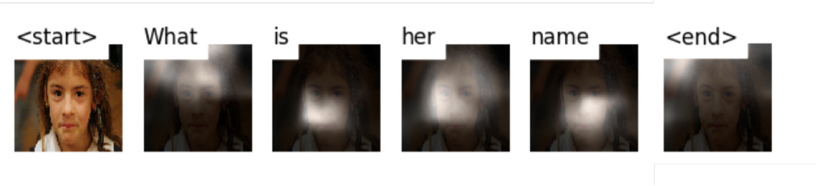


Figure 3.7: Attention mechanism.

The questions are generated in the following way through a beam search mechanism [33]. A set of word candidates is generated by the LSTM at each decoding step. We can see an example of a beam search using a size of 2 at Figure 3.8. The generated question always begins with <start> and ends up with <end>. To chose the best questions to ask per image, at the first decode step, N word candidates are generated, each one with an associated score. In the second step, N words are generated for each of these N first words and [first word, second word] combinations with higher additive scores are chosen. For each of the selected second words, N more word candidates are generated and the best combinations are selected. This is repeated for each decode step until the word <end> is generated in every sequence. Finally the sequences with the highest score are the selected questions. Beam search tries to approximate $y = argmax_{\hat{y}}\ p(\hat{y}|I)$. We tested different beam sizes and we achieve the best performance in terms of quality and computation-speed with a beam size of 7.

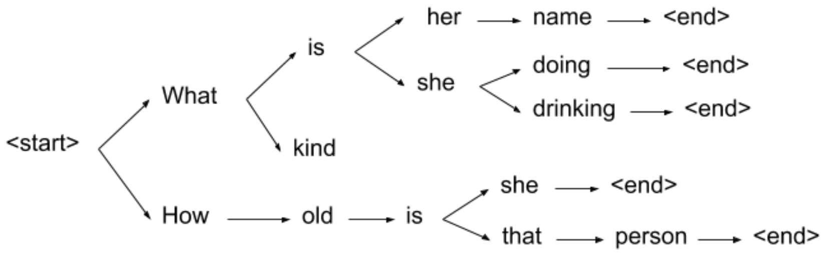


Figure 3.8: Beam search mechanism.

3.5 Chatbot Network

The core of our chatbot model is a sequence-to-sequence [30]. This architecture uses a RNN to encode a variable-length sequence to obtain a large fixed dimensional vector representation and another RNN to decode the vector into a variable-length sequence.

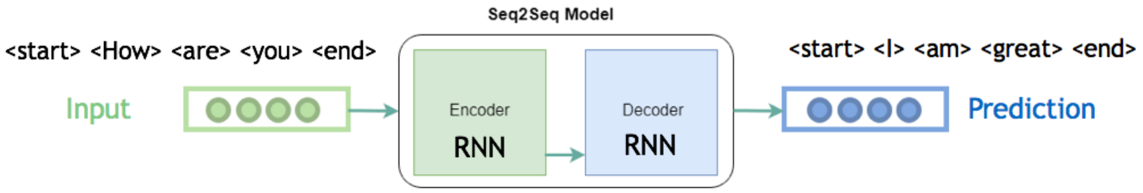

Seq2Seq Model
<start> <How> <are> <you> <end>
Input
Encoder
RNN
Decoder
RNN
<start> <I> <am> <great> <end>
Prediction

Figure 3.9: Sequence-to-Sequence architecture.

The encoder iterates through the input sentence one word at each time step producing an output vector and a hidden state vector. The hidden state vector is passed to the next time step, while the output vector is stored. We use a bidirectional LSTM, meaning we use 2 LSTMs one fed in sequential order and another one fed in reverse order. The outputs of both networks are summed at each time step, so we encode past and future context.

The final hidden state h_t^{enc} is fed into the decoder as the initial state h_0^{dec}. By using an attention mechanism, the decoder uses the encoder's context vectors, and internal hidden states to generate the next word in the sequence. It continues generating words until it outputs an $<$end$>$ token, representing the end of the sentence. We use an attention layer to multiply attention weights to encoder's outputs to focus on the relevant information when decoding the sequence. This approach have shown better performance on sequence-to-sequence models [2].

3.6 Training Procedure

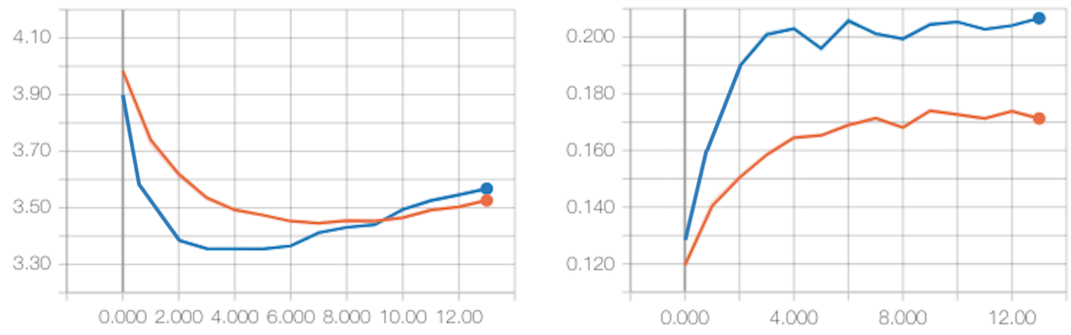
We trained the VQG and Chatbot networks independently by minimizing the Cross-entropy loss, which measures the performance of a classification model with C classes whose output is a probability value.

$$L = \sum_{C=1}^{M} y_{o,c} \log(p_{o,c}) \tag{3.2}$$

Where M is the number of classes, y is a binary indicator (0 or 1) if class label is the correct classification for observation o, and p is the predicted probability observation that o is of class c.

Both models were trained using Stochastic Gradient Descent with ADAM optimization [14] and a learning rate of 1e-4. We used Dropout regularization [28] which prevents from over-fitting by dropping some units of the network. We used BLEU metric on the validation set for model selection, which is a measure of similitude between generated and target sequences of words. BLEU has been the most commonly used metric so far in the image description literature [21]. Both models took 10 days to train on the GPU.

In the following figures we can see the values of Cross-Entropy loss (3.11a) and BLEU (3.11b) per epoch for the same model trained with different datasets. The blue line is the VQG model using a maximum question length of 6 words, while the orange line is the same model using a maximum length of 12 words. An epoch is when the entire dataset is passed forward and backward through the neural network. We divide the data in batches of 32 and the model, during training, updates the weights using Gradient Descent optimization algorithm to generate, at each time step, a better prediction. We can see that the validation loss decreases up to a point and then (at around 3.4) starts to increase. We can see that the loss starts to increase at different times for each of the models. BLEU metric instead, increases up to a point and then it remains stable with little variations. This means the predictions are getting closer to the target until at a point where the model starts to overfit the training set. We stop training our model when we obtain the highest BLEU, which is at epoch 13th. It is interesting to see that the model that uses more data (all sequences up to 12 words) has a lower BLUE score, this is probably because it is more difficult for it to learn to generate questions similar to its training set. Thus, we use the model which generates questions up to 6 words in our test with patients.



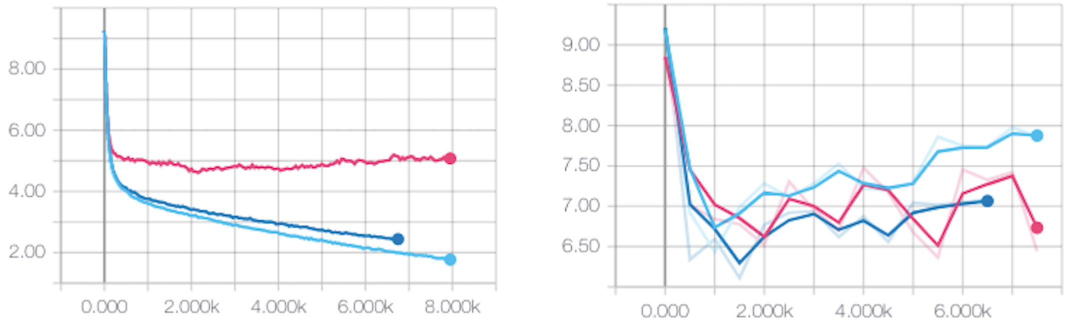
(a) Validation loss

(b) BLEU metric

Figure 3.10: Validation curves of VQG model

The VQG encoder is composed of 2048 neuron cells. This means a tensor of this size is passed to the decoder, which generates a sequence of word using this information. The VQG decoder has an attention layer of 512 followed by an embedding layer of 512 and a LSTM with the same size. We use a dropout of 50% and a beam search of 7 for decoding. The vocabulary we use consists of all words seen 3 or more times in the training set, which amounts to 11.214 unique tokens. Unknown words are mapped to to an <unk> token during training, but we do not allow the decoder to produce this token at test time.

In the Chatbot model we use a hidden size of 500 and Dropout regularization of 25%. We first train it with Persona-chat and then fine-tune it with Cornell dataset as we saw the responses of the chatbot improved. For the hyperparameter setting, we tried several models changing the hyperparameters (batch size, learning rate, neural cells dimension...) and we chose the ones that decrease the most our training and validation loss. In the following figure we can see the curves for three of the chatbot models we trained. We chose the model that has the dark blue loss as it is the one with a lower validation loss, which corresponds to the hyperparameters commented before.



(a) Training loss

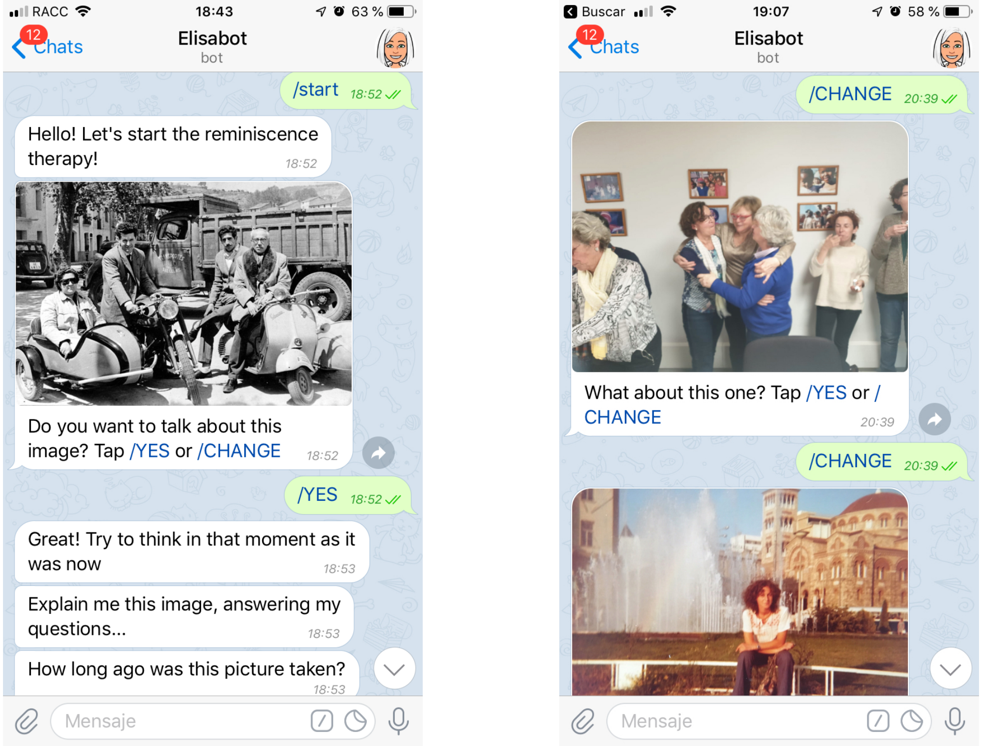
(b) Validation loss

Figure 3.11: Loss of Chatbot model

3.7 Telegram User Interface

The user interface of Elisabot is Telegram, an instant messaging application available for smartphones or computers. We selected it because it easy to use and it offers an API for developers to connect bots to the Telegram system. It enables to create special accounts for bots which do not require a phone number to set up.

Telegram is only the interface for the code running in the server. The bot is executed via an HTTP-request to the API. Users can start a conversation with Elisabot by typing @TherapistElisabot in the searcher and executing the command `/start`, as can be seen in the Figure 3.12. Messages, commands and requests sent by users are passed to the software running on the server. We add `/change`, `/yes` and `/exit` commands to enable more functionalities. `Change` changes the image in case the user does not want to talk about it because he/she does not remember it, `/yes` accepts the image which is going to talk about and `/exit` finishes the dialogue with Elisabot. The commands can be executed either by tapping on the linked text or typing them.



RACC 18:43 63 %
Chats 12
Elisabot
bot

/start 18:52

Hello! Let's start the reminiscence therapy! 18:52

Do you want to talk about this image? Tap /YES or /CHANGE 18:52

/YES 18:52

Great! Try to think in that moment as it was now 18:53

Explain me this image, answering my questions... 18:53

How long ago was this picture taken? 18:53

Mensaje

(a) /Start

Buscar 19:07 58 %
Chats 12
Elisabot
bot

/CHANGE 20:39

What about this one? Tap /YES or /CHANGE 20:39

/CHANGE 20:39

Mensaje

(b) /Change

Figure 3.12: Elisabot running on Telegram application

Chapter 4

Experimental Results

In this section, we describe the experiments we do with our models, we show some qualitative results of the interaction with Elisabot, and we present a user study with real patients.

4.1 Qualitative results

Our first goal was to generate meaningful questions about the provided images. In the following Table we can see the generated questions by our VQG model for each of the proposed images. These images were not in the training set as they were taken from my family photo albums, so the model was the first time that had these images as input. By looking at the questions, we can easily realise that the model is not only able to generate questions grammatically correct, but to understand the content of the image and ask about it. Furthermore, we not only wanted to generate only a question per image, but to bring out up to five questions for a deeper conversation.

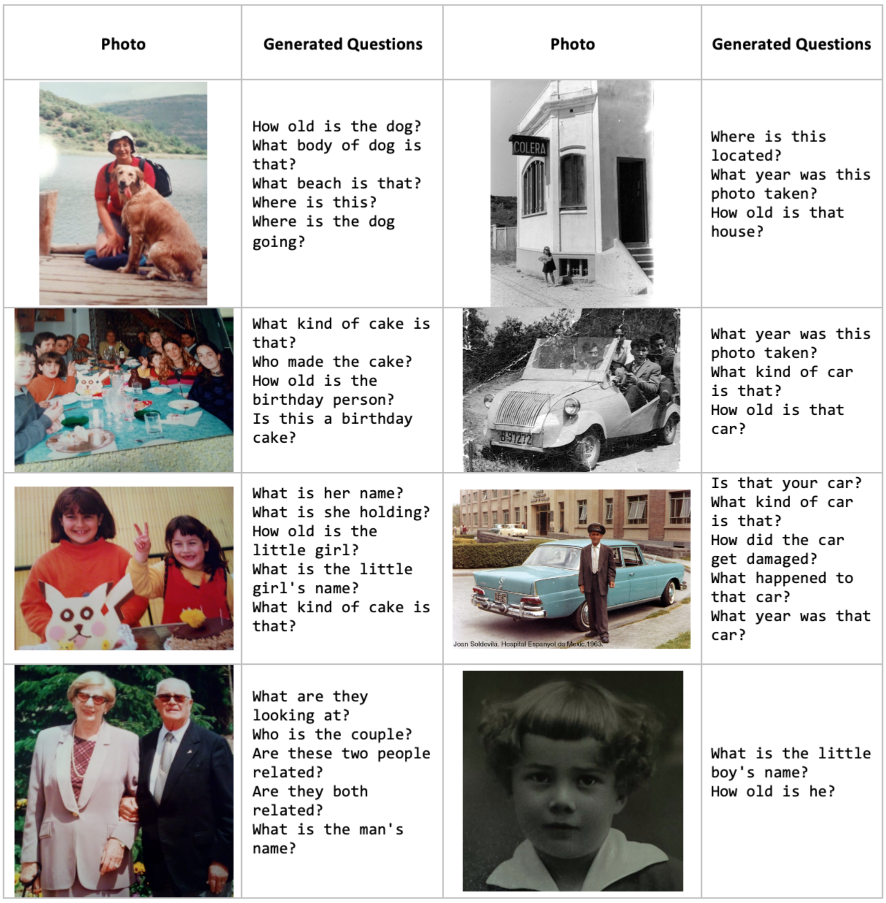


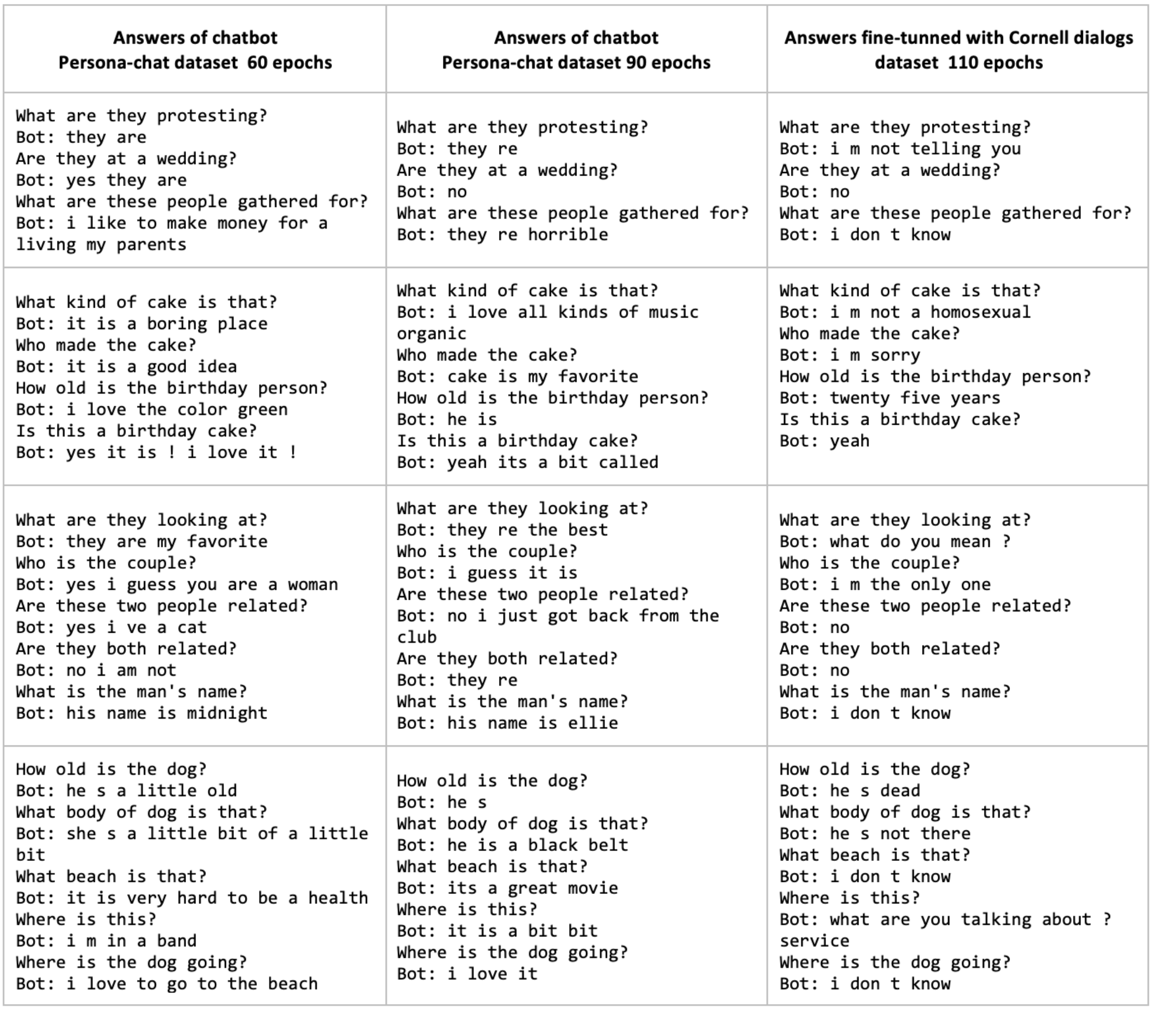
Photo	Generated Questions	Photo	Generated Questions
	How old is the dog? What body of dog is that? What beach is that? Where is this? Where is the dog going?		Where is this located? What year was this photo taken? How old is that house?
	What kind of cake is that? Who made the cake? How old is the birthday person? Is this a birthday cake?		What year was this photo taken? What kind of car is that? How old is that car?
	What is her name? What is she holding? How old is the little girl? What is the little girl's name? What kind of cake is that?		Is that your car? What kind of car is that? How did the car get damaged? What happened to that car? What year was that car?
	What are they looking at? Who is the couple? Are these two people related? Are they both related? What is the man's name?		What is the little boy's name? How old is he?

Figure 4.1: Generated questions

We used different kind of pictures as we wanted a robust model able to work with all kind of images. In the Figure 7.1 we can see old black and white photos and colorful photos, both with vertical and horizontal shapes. Most of the pictures show people because these are the type of pictures used in reminiscence therapy as they evoke memories from the relationships with that persons. We also tested pictures showing pets, cars and motorbikes as some people are very attached to them, triggering strong emotions.

Most of the generated questions could be asked by a person. Nonetheless, some of the questions such as *How did the car get damaged?* do not correspond with the image's content, or are not well formulated like *What body of dog is that?*, but can be understood.

About the chatbot module, we first evaluate it by passing the generated questions from VQG to see what kind of answers can be generated and then we test it by having a conversation with it. In the following Table we can see the evolution of the generated answers as we trained the model. The first two columns are the same model trained with Persona-chat dataset for different periods of time. The first one is trained for 60 epochs, while the second one is trained for 90 epochs. We can see a slight improvement in the relation between the question and the answer due to the training time. The third model is trained for more time and fine-tuned with the Cornell-movies dataset.



Answers of chatbot Persona-chat dataset 60 epochs	Answers of chatbot Persona-chat dataset 90 epochs	Answers fine-tunned with Cornell dialogs dataset 110 epochs
What are they protesting? Bot: they are Are they at a wedding? Bot: yes they are What are these people gathered for? Bot: i like to make money for a living my parents	What are they protesting? Bot: they re Are they at a wedding? Bot: no What are these people gathered for? Bot: they re horrible	What are they protesting? Bot: i m not telling you Are they at a wedding? Bot: no What are these people gathered for? Bot: i don t know
What kind of cake is that? Bot: it is a boring place Who made the cake? Bot: it is a good idea How old is the birthday person? Bot: i love the color green Is this a birthday cake? Bot: yes it is ! i love it !	What kind of cake is that? Bot: i love all kinds of music organic Who made the cake? Bot: cake is my favorite How old is the birthday person? Bot: he is Is this a birthday cake? Bot: yeah its a bit called	What kind of cake is that? Bot: i m not a homosexual Who made the cake? Bot: i m sorry How old is the birthday person? Bot: twenty five years Is this a birthday cake? Bot: yeah
What are they looking at? Bot: they are my favorite Who is the couple? Bot: yes i guess you are a woman Are these two people related? Bot: yes i ve a cat Are they both related? Bot: no i am not What is the man's name? Bot: his name is midnight	What are they looking at? Bot: they re the best Who is the couple? Bot: i guess it is Are these two people related? Bot: no i just got back from the club Are they both related? Bot: they re What is the man's name? Bot: his name is ellie	What are they looking at? Bot: what do you mean ? Who is the couple? Bot: i m the only one Are these two people related? Bot: no Are they both related? Bot: no What is the man's name? Bot: i don t know
How old is the dog? Bot: he s a little old What body of dog is that? Bot: she s a little bit of a little bit What beach is that? Bot: it is very hard to be a health Where is this? Bot: i m in a band Where is the dog going? Bot: i love to go to the beach	How old is the dog? Bot: he s What body of dog is that? Bot: he is a black belt What beach is that? Bot: its a great movie Where is this? Bot: it is a bit bit Where is the dog going? Bot: i love it	How old is the dog? Bot: he s dead What body of dog is that? Bot: he s not there What beach is that? Bot: i don t know Where is this? Bot: what are you talking about ? service Where is the dog going? Bot: i don t know

Figure 4.2: Generated answers to the generated questions

The fine-tuned model is clearly the best one, it has learned to generate sentences with better grammatical structure and more coherent answers, although it has a tendency to produce non-specific answers like *I don't know*, which is a common problem for current state-of-the-art conversational chatbots.

In the following lines we can see two samples of a conversation with our chatbot model, where each sentence starting with ”>” is an input to our seq2seq model.In this first sample, the chatbot model is clearly overfitted as it repeats the same sentence twice. While in the second sample we can see that the model is able to follow the conversation.

Overfitted chatbot model:

```
> hi
Elisabot:  hello !  how are you ?
> fine and you?
Elisabot :  i am good do you have any hobbies ?
> yes i like to ski
Elisabot:  i love to play sports .  it is my favorite
> I have a sister
Elisabot :  i have a dog .  i love them .
> I have a dog too
Elisabot:  i have a dog .  i love them
> ok
Elisabot :  i am not really a good person
> why?
Elisabot :  i love to play sports .  it is my favorite
```

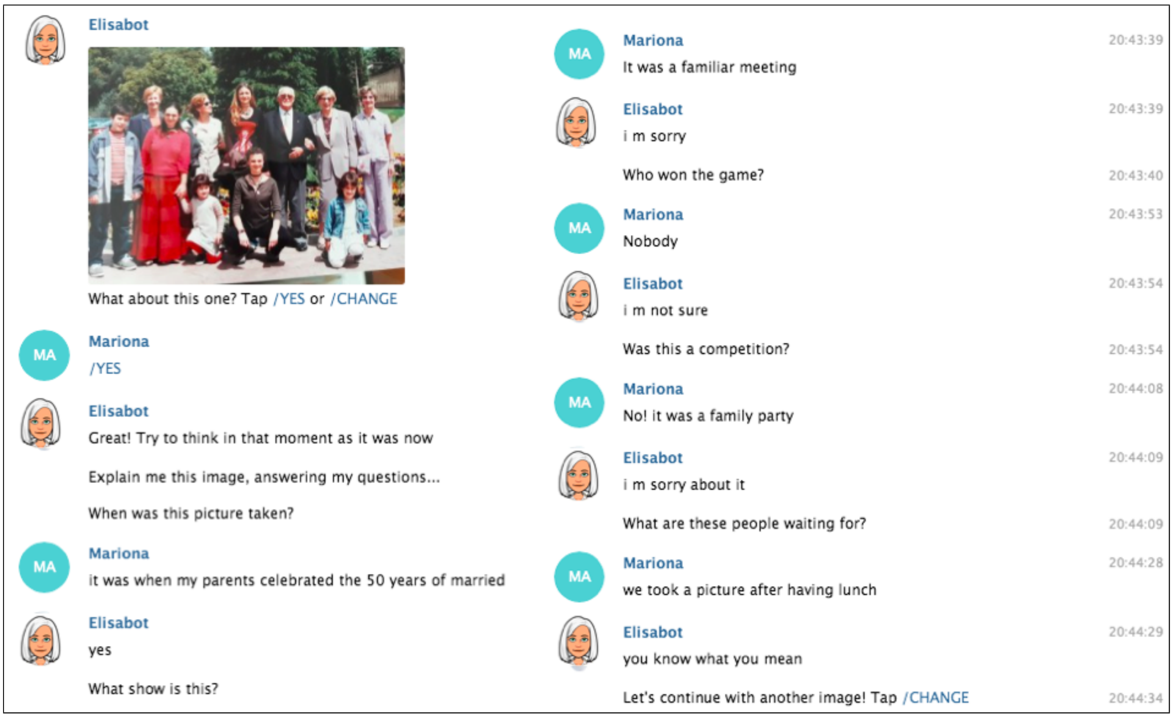
Final model:

```
> hello
Elisabot:  hello how are you ?
> fine and you?
Elisabot :  i m great !
> I have a sister, do you have any brother or sister?
Elisabot :  i have a iphone but i have no idea .
> ok
Elisabot :  i am a pro fan of thrones
> me too
Elisabot :  i like to eat healthy
> have you seen the last chapter?
Elisabot :  no i have not .
> where do you live?
Elisabot :  i live in the mountains
```

4.2 User Study

We designed a session study where users with and without mild cognitive impairment interacted with the system with the help of a doctor and myself. The purpose was to study the acceptability and feasibility of the system with patients of mild cognitive impairment. The users were all older than 60 years old. The sessions lasted 30 minutes and were realized by using a laptop computer with Telegram. As Elisabot's language is English I translated the questions to the users and the answers to Elisabot.

We first tested Elisabot with a user without mild cognitive impairment to check if it was easy for her to do the therapy using Elisabot. In the following Figure 4.3 we can see a sample of the conversation.



Elisabot
What about this one? Tap /YES or /CHANGE

Mariona
/YES

Elisabot
Great! Try to think in that moment as it was now

Explain me this image, answering my questions...

When was this picture taken?

Mariona
it was when my parents celebrated the 50 years of married

Elisabot
yes

What show is this?

Mariona 20:43:39
It was a familiar meeting

Elisabot 20:43:39
i m sorry

Who won the game? 20:43:40

Mariona 20:43:53
Nobody

Elisabot 20:43:54
i m not sure

Was this a competition? 20:43:54

Mariona 20:44:08
No! it was a family party

Elisabot 20:44:09
i m sorry about it

What are these people waiting for? 20:44:09

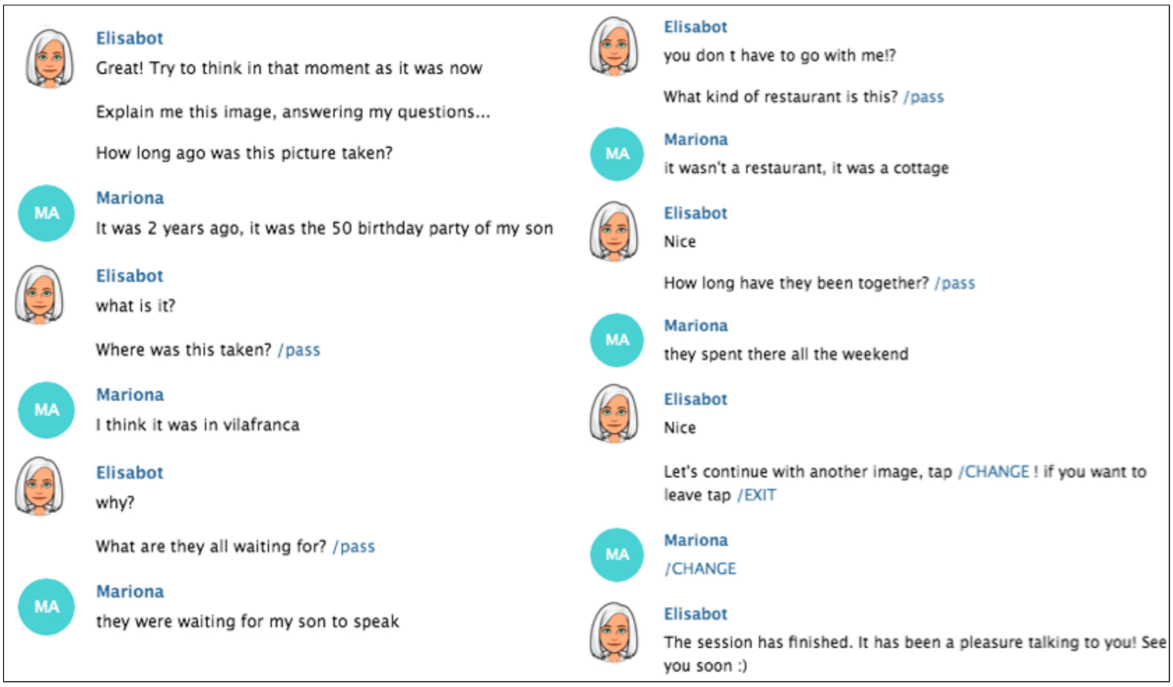
Mariona 20:44:28
we took a picture after having lunch

Elisabot 20:44:29
you know what you mean

Let's continue with another image! Tap /CHANGE 20:44:34

Figure 4.3: Session study with user without cognitive impairment

Then, we tested it with patients from *Hospital Consorci de Terrassa*. A sample of these sessions is provided in Figure 4.4, without including the picture for data privacy.



Elisabot
Great! Try to think in that moment as it was now

Explain me this image, answering my questions...

How long ago was this picture taken?

Mariona
It was 2 years ago, it was the 50 birthday party of my son

Elisabot
what is it?

Where was this taken? /pass

Mariona
I think it was in vilafranca

Elisabot
why?

What are they all waiting for? /pass

Mariona
they were waiting for my son to speak

Elisabot
you don t have to go with me!?

What kind of restaurant is this? /pass

Mariona
it wasn't a restaurant, it was a cottage

Elisabot
Nice

How long have they been together? /pass

Mariona
they spent there all the weekend

Elisabot
Nice

Let's continue with another image, tap /CHANGE ! if you want to leave tap /EXIT

Mariona
/CHANGE

Elisabot
The session has finished. It has been a pleasure talking to you! See you soon :)

Figure 4.4: Session study with user diagnosed of cognitive impairment

In the first case we can see that Elisabot gets the wrong information about the image, as it keeps asking for a game or a competition. However the user enjoyed the conversation as she found it amusing and the answers and questions from Elisabot were well formulated and fast; as can bee seen on the message's timestamps. Moreover it seems that Elisabot apologized for asking wrong questions.

In the second sample all the generated questions were right according to the image content (which is not shown here for privacy), but the feedback was wrong for some answers. We can see that these was the last picture of the session as when Elisabot asks if the user wants to continue or leave, and he decides to continue, Elisabot finishes the session as there are not more pictures remaining to talk about.

At the end of the session, we administrated the survey, that can be found in appendices, to ask participants the following questions about their assessment of Elisabot:

- Did you like it?
- Did you find it engaging?
- How difficult have you found it?

Responses were given on a five-point scale ranging from *strongly disagree* (1) to *strongly agree* (5) and *very easy* (1) to *very difficult* (5). The results were 4.6 for amusing and engaging and 2.6 for difficulty. The healthy user found it very easy to use (1/5) and even a bit silly because of some of the generated questions and answers. Nevertheless, users with mild cognitive impairment found it engaging (5/5), but challenging (4/5) because of the effort they had to make to remember the answers for some of the generated questions. All the users had in common that they enjoyed doing the therapy with Elisabot.

Chapter 5

Budget

The hardware resources needed for the project were a Macintosh laptop and a NVIDIA GPU. The GPU was used during 2 months for the development of the model, which adds to 1.344 hours of computation. We compute the computation cost based on Amazon Web Services (AWS) rates[1] for *p2.xlarge* instances with one NVIDIA K80. Regarding software, we used *Pycharm Professional* which licence cost 89 € .

The main costs of this projects comes from the salary of the researches and the time spent in it. The team for the development of this thesis is formed by two senior engineers as the advisors and myself as a junior engineer. The length of the project was 28 weeks, as presented in the Gantt diagram. Assuming a commitment of 20 weekly hours and that each advisor spent an average of 1h per week on meetings, the complete costs for the project are the following:

	Amount	**Cost/hour**	**Time**	**Total**
GPU *p2.xlarge*	1	0,90 €	1.344h	1.210 €
Pycharm Professional	1	-	-	89 €
Junior engineer	1	10,00 €	700h	7.000 €
Senior engineer	2	30,00 €	28h	1.680 €
Other equipment	-	-	-	3.000 €
			Total	12.979 €

Table 5.1: Cost of the project. *Other equipment* includes campus services and employed laptop.

[1] https://aws.amazon.com/ec2/instance-types/p2/?nc1=h_ls

Chapter 6

Conclusion and Future Work

We presented a dialogue system for handling sessions of 30 minutes of reminiscence therapy. Elisabot, our conversational agent leads the therapy by showing a picture and generating some questions. The goal of the system is to improve users mood and stimulate their memory and communication skills. Two models were proposed to generate the dialogue system for the reminiscence therapy. A visual question generator composed of a CNN and a LSTM with Attention and a sequence-to-sequence model to generate feedback on the user's answers. We realize that fine-tuning our chatbot model with another data set improved the generated dialogue.

The manual evaluation shows that our model can generate questions and feedback well formulated grammatically, but in some occasions not appropriate in content. Furthermore, it has tendency to produce non-specific answers and to loss its consistency in the comments with respect to what it has said before.

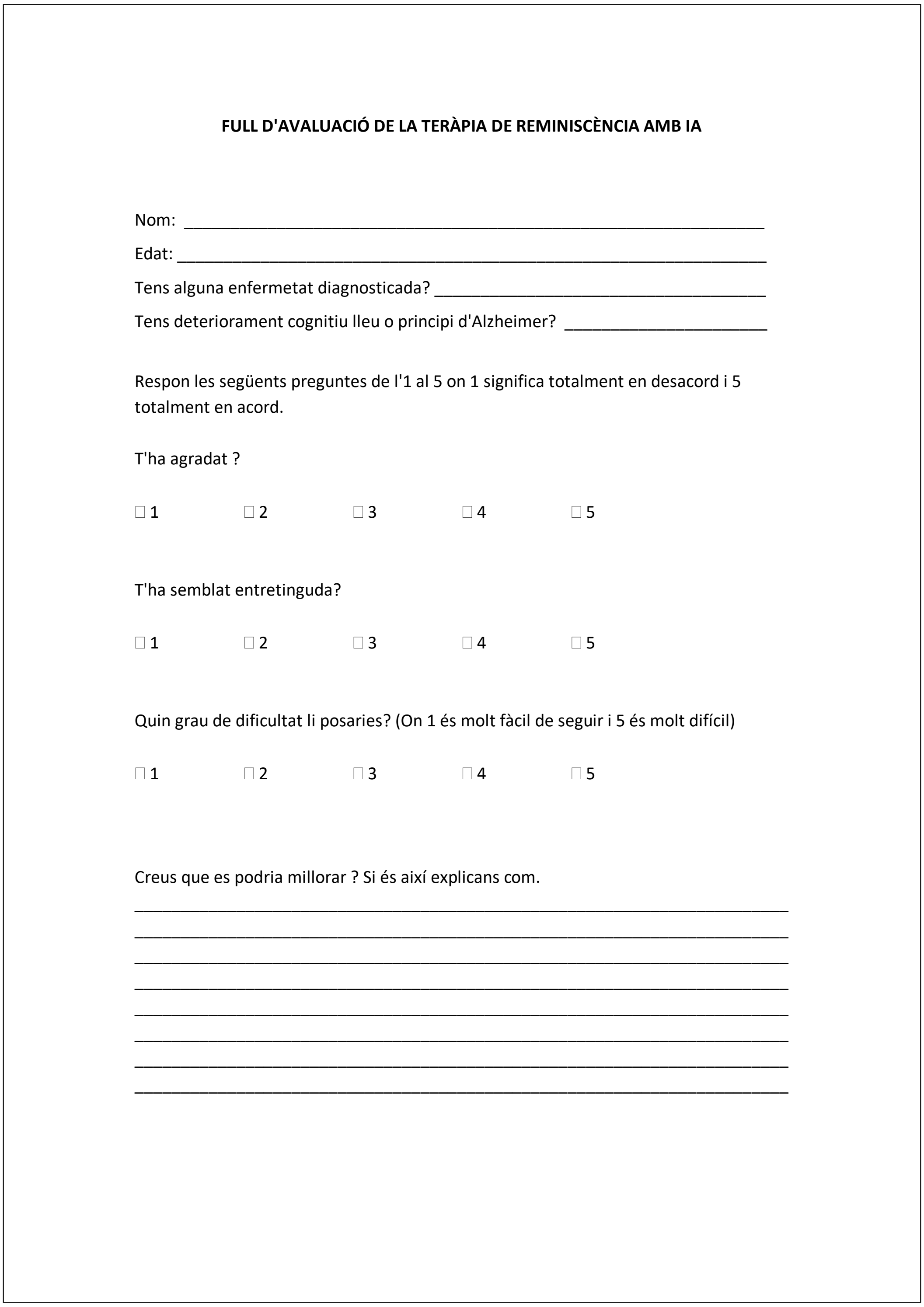
We ran a user study including three participants interacting with the dialogue system. The overall usability evaluation of the system by users with mild cognitive impairment shows that they found the session very entertaining and challenging. They had to make an effort to remember the answers for some of the questions, but they were very happy when they achieved it. Though, we see that for the proper performance of the therapy is essential a person to support the user to help him/her to remember.

This project has many possible future lines. In our future work, we suggest to train the model including the Reddit dataset which could improve a lot the chatbot model as it has many open-domain conversations. Moreover, we would like to include speech recognition and generation, as well as real-time text translation, to make Elisabot more autonomous and open to older adults with reading and writing difficulties. Furthermore, the lack of consistency in the dialogue might be avoided by improving the architecture including information about passed conversation into the model. We also think it would be a good idea to recognize feelings from the user's answers and give a feedback according to them.

Code and models are publicly available at https://github.com/marionacaros/generative-dialogue-system-for-reminiscence-telegram.

Chapter 7

Appendices



FULL D'AVALUACIÓ DE LA TERÀPIA DE REMINISCÈNCIA AMB IA

Nom: ____

Edat: ____

Tens alguna enfermetat diagnosticada? ___________________________________

Tens deteriorament cognitiu lleu o principi d'Alzheimer? ______________________

Respon les següents preguntes de l'1 al 5 on 1 significa totalment en desacord i 5 totalment en acord.

T'ha agradat ?

□ 1 □ 2 □ 3 □ 4 □ 5

T'ha semblat entretinguda?

□ 1 □ 2 □ 3 □ 4 □ 5

Quin grau de dificultat li posaries? (On 1 és molt fàcil de seguir i 5 és molt difícil)

□ 1 □ 2 □ 3 □ 4 □ 5

Creus que es podria millorar ? Si és així explicans com.

____

____

____

____

____

____

____

____

Figure 7.1: Evaluation survey for the session with Elisabot

Bibliography

[1] Association Alzheimer's. 2015 alzheimer's disease facts and figures. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 11(3):332, 2015.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] Antoine Bordes, Y-Lan Boureau, and Jason Weston. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*, 2016.

[4] Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*, 2011.

[5] Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, 2014.

[6] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[7] Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. Approximating interactive human evaluation with self-play for open-domain dialog systems. *arXiv preprint arXiv:1906.09308*, 2019.

[8] Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. The photobook dataset: Building common ground through visually-grounded dialogue. *arXiv preprint arXiv:1906.01530*, 2019.

[9] Richard A Hickman, Arline Faustin, and Thomas Wisniewski. Alzheimer disease and its growing epidemic: risk factors, biomarkers, and the urgent need for therapeutics. *Neurologic clinics*, 34(4):941–953, 2016.

[10] Alina Huldtgren, Anja Vormann, and Christian Geiger. Reminiscence map: Insights to design for people with dementia from a tangible prototype. 2015.

[11] Natasha Jaques, Asma Ghandeharioun, Judy Shen, Craig Ferguson, Noah Jones, Agata Lapedriza, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:*.

[12] Zongcheng Ji, Zhengdong Lu, and Hang Li. An information retrieval approach to short text conversation. *arXiv preprint arXiv:1408.6988*, 2014.

[13] Claudia H Kawas and Maria M Corrada. Alzheimer's and dementia in the oldest-old: a century of challenges. *Current Alzheimer Research*, 3(5):411–419, 2006.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Stefan Kopp, Mara Brandt, Hendrik Buschmeier, Katharina Cyra, Farina Freigang, Nicole Krämer, Franz Kummert, Christiane Opfermann, Karola Pitsch, Lars Schillingmann, et al. Conversational assistants for elderly users–the importance of socially cooperative dialogue. In *AAMAS Workshop on Intelligent Conversation Agents in Home and Geriatric Care Applications*, 2018.

[16] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[17] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[19] Jose Luis Monteagudo. Capacidades y oportunidades de innovación en tic para alzheimer. 2013.

[20] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Larry Zitnick, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. *CoRR*, abs/1603.06059, 2016.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[22] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[23] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

[24] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[25] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[26] Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. Engaging image chat: Modeling personality in grounded dialogue. *arXiv preprint arXiv:1811.00945*, 2018.

[27] Maite Solas, Elena Puerta, and Maria J Ramirez. Treatment options in alzheimer s disease: The gaba story. *Current pharmaceutical design*, 21(34):4960–4971, 2015.

[28] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[29] Ponnusamy Subramaniam and Bob Woods. The impact of individual reminiscence therapy for people with dementia: systematic review. *Expert Review of Neurotherapeutics*, 12(5):545–555, 2012.

[30] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[31] Christiana Tsiourti, Maher Ben Moussa, João Quintas, Ben Loke, Inge Jochem, Joana Albuquerque Lopes, and Dimitri Konstantas. A virtual assistive companion for older adults: design implications for a real-world application. In *Proceedings of SAI Intelligent Systems Conference*, pages 1014–1033. Springer, 2016.

[32] Dina Utami, Timothy Bickmore, Asimina Nikolopoulou, and Michael Paasche-Orlow. Talk about death: End of life planning with a virtual agent. In *International Conference on Intelligent Virtual Agents*, pages 441–450. Springer, 2017.

[33] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

[34] Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[35] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[36] Fei-Yue Wang, Jun Jason Zhang, Xinhu Zheng, Xiao Wang, Yong Yuan, Xiaoxiao Dai, Jie Zhang, and Liuqing Yang. Where does alphago go: From church-turing thesis to alphago thesis and beyond. *IEEE/CAA Journal of Automatica Sinica*, 3(2):113–120, 2016.

[37] Joseph Weizenbaum et al. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[38] Paul J Werbos et al. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[39] Bob Woods, Laura O'Philbin, Emma M Farrell, Aimee E Spector, and Martin Orrell. Reminiscence therapy for dementia. *Cochrane database of systematic reviews*, (3), 2018.

[40] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*, 2016.

[41] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[42] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.

[43] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629, 2018.